DOCUMENT RESUME

ED 152 693 95 SP 012 230

AUTHOR Borich, Gary D.
TITLE Three School Based Models for Conducting Follow-up Studies of Teacher Education and Training.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 77
GRANT OB-NIE-C-78-0116
NOTE 75p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Case Studies; Comparative Analysis; *Evaluation Methods; *Followup Studies; Fundamental Concepts; *Models; *Program Evaluation; *Research Methodology; *Teacher Education

ABSTRACT

The rise of interest in the evaluation of teacher education and training during the decade 1967-1977 is charted; a review of related concepts and studies is presented; and three evaluation models for conducting follow-up studies on training effectiveness are examined. Three issues arising to prominence in the last decade are identified and discussed: pay-off evaluation, process evaluation, and goal assessment. A review of the status of process-product findings and studies that produced them are linked with the concepts of competency-based evaluation and follow-up evaluation. Five major process-product studies, comprising the research base from which many teacher competencies have been devised, are examined, their general methodological concepts summarized, and findings listed. Common characteristics of these product-process investigations are correlated, and limitations of these methodologies in the conduct of inservice evaluations are discussed. Three generic models are presented to portray the variety of methodologies commonly employed in teacher evaluations: Needs Assessment, Relative Gain, and Process-Product. Various context in which these three models can be used are discussed, and case studies are presented to illustrate the contextual characteristics most often associated with each of the evaluation models. Key variables that suggest the use of one model over another are also identified. (HJB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AACTE

ED152693

Three School Based Models for Conducting Follow-up Studies of Teacher Education and Training

Gary D. Borich
College of Education
The University of Texas at Austin

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Portions of this paper were prepared for the organization for Economic Co-operation and Development (OECD), Center for Educational Research and Innovation, Paris, and are based on the author's book The Appraisal of Teaching: Concepts and Process. Reading, MA: Addison-Wesley, 1977.

This paper was prepared under the auspices of the Research and Development Center for Teacher Education, The University of Texas at Austin, NIE Grant No. OB-NIE-C-78-0116.

SP012 230

THREE SCHOOL BASED MODELS FOR CONDUCTING FOLLOW UP STUDIES OF TEACHER EDUCATION AND TRAINING

Gary D. Borich

The University of Texas at Austin

It is unlikely that any educator in the United States has failed to notice that teacher education in the last decade has been wrought with controversy, anxiety, and potential. It is as if all the emotion associated with differing viewpoints and beliefs had been subdued for decades and then unleashed between 1967 and 1977. As a result, various teacher education positions, movements, and "camps" have suddenly surfaced, producing side effects such as increased citizen interest in the public schools, the squaring off of teachers and school administrators on the issue of performance evaluation, and an almost phenomenal accumulation of books, articles, and symposia dealing with the process of schooling.

This period of change and unrest has also been marked by widespread and wide ranging discussions in and about the field of evaluation. A fledgling discipline a decade earlier, evaluation has been written about and discussed in some form or another at almost every national gathering of educators and social scientists in the last 10 years. During this period developments in the broader field of evaluation began to influence the narrower discipline of teacher evaluation, particularly with regard to the methodologies used to assess teacher training programs and the performance of teachers.

Three issues in the field of evaluation rose to prominence during this period and are now beginning to make an impact on the evaluation of teacher education and training. These are the concepts of pay-off evaluation, process evaluation, and goal assessment.

Pay-Off Evaluation

First brought to prominence by Scriven in a monograph entitled The Methodology of Evaluation (1967), the concept of pay-off evaluation has achieved considerable acceptance among evaluators. The essence of "pay-off" evaluation is its capacity to ascertain the ultimate value of a program, product, or procedure by comparing it to some realistic alternative program, product, or procedure. Hence, the true merit of the object or process being evaluated is determined by examining its effectiveness vis-a-vis some other object or process having the same or a similar purpose. The utility of pay-off evaluation depends on the nature of the comparison chosen. If, for example, a new third-grade reading text were being evaluated, a comparison consisting of a no-treatment control would probably be less informative than a comparison involving some alternative form of instruction. Simply stated, the tougher the competition, the more credible the pay-off evaluation.

The word "pay-off," however, also implies finality and reward; and "pay-off" evaluation can indeed address the ultimate concern of the program or product developer. If this concern is to produce a better product, then the evaluator must ask "better than what?" If the evaluator selects an evaluation procedure that establishes the efficacy of the product above all reasonable competition, he or she has responded to this most important question. If the program or product is "effective," there is a pay-off or reward for the population served by that program or product, usually in the form of knowledge gained, skills learned, or performance achieved. Thus, measurement of the program's impact on the population being served is often considered the sine qua non of pay-off evaluation.

How has the concept of pay-off evaluation influenced the evaluation of teacher education and training? This question can be answered most directly by comparing the teacher research and evaluation literature before and after the inception of "pay-off" evaluation. One can gauge the relative recency of this concept by examining a landmark study by Ryans published in 1960. From 1954 to approximately 1958, Ryans evaluated the characteristics of teachers in the U.S., in the largest and most comprehensive study of its kind to that date. This study has been and continues to be frequently cited in major articles, books, and monographs on teacher behavior and evaluation--a fact that attests to its use and acceptance by other evaluators and researchers. Yet very little, if any, pay-off evaluation appears in this book-length study. If "pay-off" evaluation is defined simply as a comparison between groups, then Ryans may be credited with using it, since he at one point divides a small portion of his sample into 3 groups on the basis of observed characteristics and then compares the performance of these groups to their self-reported teaching styles. While the data suggest some conclusions about the desirability of various teaching styles, their primary purpose was to cross validate data sources.

Pay-off evaluators, however, may argue that the Ryans study employs no pay-off evaluation whatsoever. If pay-off evaluation is defined not merely as the determination of relative merit, but instead as the determination of relative merit for <u>an appropriate target group</u>, then the Ryans study cannot be said to have used pay-off evaluation. Since the teacher was the focus of Ryans study, just as a program or product is the focus of an evaluation study, the appropriate population on whom the "pay-off" must be measured is the client benefitting from the teacher's activity--that is, the pupil. The

effects of various teacher characteristics on pupil behavior, affective or cognitive, were not a part of Ryans' study.

The concept of pay-off evaluation, particularly its focus on appropriate target groups, is becoming the standard against which evaluations of teacher education and training in the U.S. are being judged. These evaluations have become known as "product" studies, a label indicating that the performance of pupils, i.e., the end product of some training experience, is being measured. Such studies may or may not include an examination of the processes that are supposed to lead to product performance, but, in either case, the preponderance of pupil behavior in the design of most recent studies is unmistakable. Put another way, an evaluation of teacher education or training that examined only teacher behavior (or characteristics as did Ryans) would be rated less favorably today than a study that focused exclusively on pupil behavior (as a result of some product or program) or on the relationships between teacher and pupil behavior. For ill or good, Ryans' study, and studies like it, are being replaced largely by research that employs measures of pupil behavior as the ultimate gauge of teacher behavior or training.

Process Evaluation

A second concept that has influenced the evaluation of teacher education and training in the U.S. is process evaluation. First popularized by Stufflebeam et al. in 1971, and since incorporated into the models and writings of many educational evaluators, process evaluation concerns the manner and extent of program or product implementation. Process evaluations determine whether the program or product is being employed by the target group as its designers

and developers intended. Thus, the concept is dynamic and on-going, requiring data collection throughout the period of program or product implementation. In this respect, process evaluation differs from pay-off evaluation which requires the evaluator to wait until the program is completed before measuring its effect on the target group.

How has the concept of process evaluation influenced teacher education and training? Because process evaluation developed concurrently in two areas--teacher education and evaluation--it is difficult to ascribe its origin to either field. While its rationale is best expressed in the evaluation literature, its most concrete application, specifically its adaptation to the observation of teachers,is in the field of teacher education. The Ryans study can again be used as a case in point.

Though not described as a process study, Ryans' work is useful in clarifying the meaning of the term prior to its development and use in the field of evaluation. Process behavior prior to and during Ryans' work was a phrase commonly used to note the general demeanor, perspective, or disposition of the teacher. Terms such as "warm," "enthusiastic," "systematic," and "business-like" were typical of those used to describe the teacher's classroom behavior. These process judgments were generally made at the end of a prescribed evaluation period by the teachers themselves, by classroom observers, or by both on the basis of unspecified or vaguely defined criteria. In its early use, the term process came to stand for the atmosphere or climate in the classroom, but not the specific behaviors responsible for the atmosphere or climate that was recorded. Thus, process variables became subsets of style or personality variables which were associated with other, more specific variables that were not recorded. In short, the term process and the variables so

characterized were used as general categories under which to subsume the many specific and discrete behaviors that contributed to the overall impression recorded by the process observer.

After publication of the Ryans study, the concept of process evaluation evolved more rapidly in the field of teacher evaluation than in the broader field of evaluation, where a specific form of instrumentation has yet to be associated with it. During this period, new process instruments describing classroom behavior were appearing at a staggering pace. Known generally as "low-inference" classroom observation scales, these instruments define classroom process behavior not in the broad terms used in the past, but instead as specific and independent units positioned, usually, along several general dimensions. The well-known and almost universally disseminated Flanders Interaction Analysis Scale is an example. Differentiating direct and indirect teaching, the instrument separates classroom interaction into 10 discrete categories: Teacher Accepts Feelings, Teacher Praises or Encourages, Teacher Accepts or Uses Ideas of Student, Teacher Asks Questions, Teacher Lectures, Teacher Gives Directions, Teacher Criticizes or Justifies Authority, Student Talk-Response, Student Talk-Initiation, and Silence or Confusion. While the Flanders system is the most frequently cited process measure of this type, it by no means reflects the most specific or discrete process variables defined since the Ryans study. A host of other instruments were developed subsequent to the Flanders' scale and some of these (e.g., the Brophy-Good Dyadic Interaction System, 1970) divide classroom behaviors into as many as 150 discrete units. The generation and dissemination of these literally hundreds of low-inference classroom observation systems (Simon & Boyer, 1970) eventually led to their use (or the use of reasonable facsimiles) in the broader field of evaluation.

Process evaluation, particularly low-inference measurement, has significantly influenced the evaluation, if not the very nature, of teacher education and training. With the means to measure discrete classroom behaviors came the tendency to judge them, to regard some as more desirable than other, some more "learnable" than others, and some more relevant to pupil outcome than others. Because some low-inference process behaviors have no theoretical rationale to justify their measurement, their raison d'etre, especially in teacher training curricula, was often a matter of opinion. Thus, during some periods and at some teacher training institutions, a particularly valued observation instrument provided both the prima facie content for the training curriculum and the means by which to evaluate that training. It is probably true that at least some of the content of "innovative" teacher training programs developed during the 1960s can be traced to constructs defined by a half dozen of our most popular classroom observation systems. This trend has left a residue of sophisticated process measures, but less than adequate evidence of their validity for either training or evaluation. Consequently, the past decade of teacher research and evaluation in the U.S. has been largely influenced by the need to link newly defined teacher processes with pupil outcomes in what are called process-product studies. This research often constitutes both a post facto evaluation of process instrumentation and an attempt to find "learnable" teacher behaviors that can account for pupil achievement. Thus, the evaluation of teacher education and training has incorporated the notions of both process and product.

Goal Assessment

A third concept that has influenced the evaluation of teacher education and training in the U.S. is goal assessment. Although perhaps not as

prominent in the work of the teacher educator as pay-off or process evaluation, goal assessment links the widespread practice of developing teacher competencies and the often neglected need to validate them. Like the two former concepts, goal assessment can be traced to the broader field of evaluation, primarily to the work of Stake (1967) and Scriven (1967). And, just as pay-off evaluation and process description were modified to suit the requirements of teacher evaluation, so too has the concept of goal assessment been altered.

At the outset, goal assessment was a remarkably popular concept because it stood in contradistinction to the more static definitions of evaluation that preceded it. These early definitions had equated evaluation either with measurement, or, more specifically, with measurement of program or product objectives. This measurement emphasis focused the work of the evaluator on three routine steps: (1) identifying objectives; (2) selecting instruments; and (3) quantifying differences between achieved and expected outcomes. From this conception, evaluators gained the notion that the objectives themselves were sacrosanct, unalternable and the only logical point from which to begin an evaluation. If programs had only broad goals rather than exact objectives, evaluators usually corrected such "inadequacies" at the outset by coaxing project personnel to translate these goals into specific outcomes with measurable effects. Failing this, project personnel were probed until sufficient information about program intent surfaced to allow the evaluator to compose a preliminary list of objectives. While goals and objectives were central to any evaluation during this period, their accuracy, appropriateness, and representativeness were issues neither for discussion nor measurement. The evaluator began with the assumption

that all program goals and objectives were appropriate; the task of gathering information to support or refute this assumption was beyond the scope of his work.

Stake (1967), in an article offering a somewhat different view, maintained that goals and objectives (which he termed "intents") are themselves measurable and, therefore, amenable to evaluation. Stake referred not to the effects of goals and objectives on program clients, but rather to the very nature of these goals and objectives: Are they accurately stated, appropriate to the program being evaluated, and representative of the needs or wishes of the clients to be served? In essence, Stake posited an evaluation model wherein it is not only legitimate but also mandatory that the evaluator examine the goals of various constituent groups served by the program. He used the term "logical contingency" to denote the extent to which these goals are embodied in the instructional activities of the program.

Scriven (1967), who also considers goal assessment within the evaluator's purview, expanded the concept to encompass the values underlying program objectives. Adamant about the importance of appraising values, Scriven not only confirmed their legitimacy as data, but also suggested that an independent evaluator be assigned to every project, operating entirely unconstrained, in order to undertake an unbiased assessment of the values and goals implicit in the program and its materials. Moreover, Scriven stressed the moral and ethical responsibility of the evaluator to bring to the attention of potential clients or participants value conflicts inherent in the program.

The concept of goal assessment has played an important, albeit subtle, role in the evaluation of teacher education and training. It has served as an intermediate step linking the two important processes of competency development and competency validation.

During the past decade an unprecedented number of teacher competencies has been formulated by state departments of education, metropolitan school districts, colleges and schools of education, and teacher researchers. But concomitant training has been provided for only a portion of these competencies. This situation--the slow or inadequate provision of training--is related to the origin of teacher competencies. In some cases classroom observation systems provided the impetus and the content for competency lists. In other cases, professional experience in the schools and fragments of developing theories were used to identify important teaching behaviors. Less frequently, competency lists were derived from empirical research that linked teacher and pupil behavior. And while research would appear a logical source from which to draw teacher competencies, its interpretation presented problems. Rosenshine's (1971) summary of process-product studies, for example, is typical of the research used as the basis for teacher competencies. His 11 "promising" teacher behaviors were sometimes construed by teacher trainers as competencies, per se, when in fact they represented much broader variables that encompassed many potential competencies and lacked the specificity needed to design training materials. Moreover, they were gleaned from the results of several studies, each of which assigned a different operational definition to the teacher process behavior under study. Thus, even when competencies were influenced by empirically conducted research, their implications, in terms of training and specific

classroom processes to be performed by the teacher, were not necessarily clear.

Thus, few "validated" competencies appear on the many competency lists already developed. Furthermore, because the lists differ in length, specificity, and content, they are less than definitive guides for training.

With development of the goal assessment concept, competencies could be viewed as malleable and responsive to client needs--with regard to both degree and kind of performance desired. If competencies are selected from existing lists and submitted for evaluation to the constituents of a training program (preservice and inservice teachers, teacher trainers, supervisors of student field experiences, state education personnel, and community school principals and administrators, for example), their adequacy can be judged prior to their use. Goal assessment allows the training agency to determine the perceived importance of a large number of behaviors, the initial selection of which may have been guided by only the most general notions of a training program and philosophy. The behaviors designated most important to various client groups then dictate the design of a training program. Ideally, these behaviors are then validated by correlating the performance of graduates of the training program with pupil outcome.

In the field of teacher education and training, then, goal assessment is used primarily to select training content, the efficacy of which should then be (but often is not) tested by relating teacher and pupil behavior in a subsequent process-product study.

Review of Related Concepts and Studies

The need to empirically establish process-product relationships has been, and remains, central to the effort to evaluate teacher education and training--and, ultimately, any kind of training. And, while many training experiences, particularly short-term courses for inservice teachers, have been implemented without the confirmation of process-product findings, it has been assumed in such cases that validation would be forthcoming. This section reviews the status of process-product findings and the nature of the studies that produced them. These studies can be related to two significant "movements" of the past decade: competency-based evaluation and follow-up evaluation.*

Competency-Based Movement

The word "competency" is an imprecise term, even to those who use it frequently. While it appears in the training literature repeatedly, its use and interpretation vary widely, and the list of synonyms for it is long. For example, terms such as "teacher behavior," "teacher variable," "teacher performance," and "teacher skill," have at one time or another and in one article or another been used interchangeably with the term "teacher competency." Perhaps because its origin may have been more political than substantive, the term has yet to take on a single and universally recognized meaning.

In the most general sense, a "competency" has come to stand for a skill, behavior, or performance expected of a trainee at the completion of training. While the term implies a criterion performance level--a cut-off point dividing those who have attained the competency from those who have not--no such

*Though known less as a movement than a methodology, follow-up studies of the type discussed here have attracted the attention of a large number of teacher trainers. The American Association of Colleges of Teacher Education (AACTE), for example, chose as a theme of its 1978 meeting "Follow-Up Studies." The dedication and motivation of its supporters is reminiscent of the competency-based movement of the early 1970s.

designation is included in the definition of a competency, as would be in a well-stated behavioral objective. The absence of expected proficiency levels in competency statements obscures the validity of competencies: How are we to know the level of execution at which the competency is most effective in producing desired pupil outcomes? The term "competency," while connoting an acceptable performance, in practice offers no more specificity of process or performance than the words "behavior," "variable," or "skill." It was this observation that led the author to conceive distinct and non-overlapping definitions for the terms "behavior," "variable," and "competency." According to these definitions, the three concepts are progressively more specific, with competencies derived from variables and behaviors and defined in terms of proficiency levels validated against pupil outcomes, as shown in Figure 1.

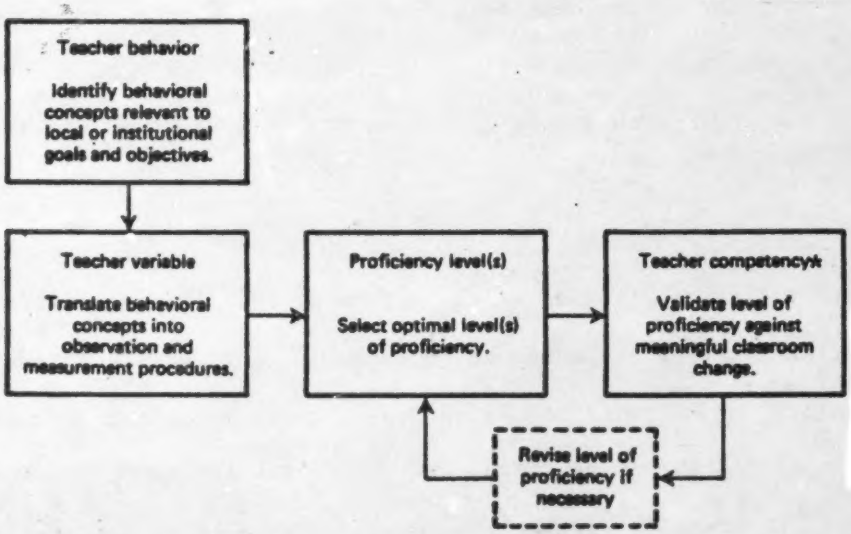

Teacher behavior
Identify behavioral concepts relevant to local or institutional goals and objectives.
Teacher variable
Translate behavioral concepts into observation and measurement procedures.
Proficiency level(s)
Select optimal level(s) of proficiency.
Teacher competency*
Validate level of proficiency against meaningful classroom change.
Revise level of proficiency if necessary

Fig. 1. The developmental task of deriving teacher competencies.

*There are three kinds of competency: knowledge competency (ability to accurately recall, paraphrase, or summarize the procedural mechanics of a behavior); performance competency (ability to accurately execute the behavior in a real or simulated environment); and, consequence competency (ability to elicit pupil learning with the behavior as recorded on tests of affective or cognitive achievement).

This conceptualization makes the term "competency" synonymous with "validated competency." That is, the word "competency" is reserved for the special case in which process-product studies have confirmed the relationship between a teaching behavior and pupil outcome. Given such process-product findings, we can estimate the optimal proficiency level or range of levels for the behavior in question.

These definitions provide a framework from which to view the contribution of individual teacher research and evaluation studies to the overall objectives of teacher education and training. Using the framework depicted in Fig. 1 for this purpose, the author found that the research is distributed in the shape of a diamond, at the apex of which are the relatively few studies that have evaluated training for the purpose of establishing proficiency levels, at the center the majority, which have evaluated training to establish relationships between teacher process and pupil product variables, and at the bottom, the relatively few studies that have evaluated training in order to determine the behaviors attained by program graduates.

Follow-Up Studies

The majority of those studies at the bottom of the diamond are "follow-up" studies. Most often, they are conducted by schools, colleges, and departments of education to determine the extent to which institutional or program objectives are being attained by graduates. They comprise rough and ready estimates of the discrepancy between the levels of competence expected of and achieved by trainees.

While the format of these studies varies, most are conducted on recent graduates of a training institution who are currently employed in the teaching

profession.* Generally, questionnaires are mailed to the graduates (or sometimes to their supervisors), but personal and telephone interviews may also be employed. These activities usually produce an accumulation of self-report data from inservice teachers, indicating the extent to which they value and apply the objectives of their training program, i.e. the "competencies" they were taught.

Although used in the formative evaluation of a training program, follow-up studies also answer the need of many institutions, state departments of education, and national or regional accreditation agencies for summative data. Program or institutional accreditation may stipulate that follow-up studies be conducted, and demands for program accountability from both the parent institution and supervising state agency can provide a major, if not the primary, impetus for follow-up. Thus, follow-up data may be used as much for confirmation as for revision of the program.

A close examination of follow-up studies, however, reveals that much of their intent is to evaluate program effectiveness vis-a-vis the competencies commonly researched in process-product studies (i.e., studies comprising the center portion of our diamond-shaped configuration). Often, the results of follow-up studies are substituted for those of more pristine field studies which observe teacher behavior and measure pupil outcome--but which may be too impractical or too costly to be conducted by a single institution. Thus, the rising popularity of follow-up studies is linked to their categorization as field studies, their relatively inexpensive format, and their acceptance as "hard" data by proponents of accountability.

*See Peer, G., & Pugues, W., "A National Survey of Teacher Education Follow-up Practices," a paper presented at the annual meeting of the American Association for Colleges of Teacher Education, February 1978, for a review of these studies.

Major Process-Product Studies

Most relevant to the objectives of this paper are studies that have attempted to identify relationships between teacher behavior and pupil outcome--those at the center of our diamond. From this research has come the justification for many of the competency statements employed in follow-up studies and other inservice evaluations of teacher education and training.

Reviewed below are five process-product studies that comprise the research base from which many teacher competencies have been derived. Though heavily funded and certainly among the most salient, these studies are not intended to represent the field. For an extensive review of other process-product research see Rosenshine (1971). This section begins with a brief overview of the five studies and concludes with a list of competency implications derived from them.

1.* Brophy, G., and Evertson, C. "Process-Product Correlations in the Texas Teacher Effectiveness Study: Final Report (RES. REP. 74-4)." Austin, Texas, Research and Development Center for Teacher Education, 1974.

This study was a 2-year effort designed to discover teacher characteristics associated with the teacher's success in producing student gains on the Metropolitan Achievement Tests (MAT). Scores on each of five MAT subtests were obtained for 3 consecutive years for the pupils of 165 2nd- and 3rd-grade teachers in an urban school system. Each student's raw mean score (grade level equivalent) was converted to a residual gain score by an adjustment that took into account the child's pretest score. These residual or adjusted gain scores were categorized by class, and a mean residual gain score on each subtest was computed for each of the 165 teachers for each of the 3 years studied. Brophy and Evertson then selected from the 165 a subsample of teachers who were notably consistent in producing achievement gains over the years, and

*References are to the original and complete research reports. Summaries of each study may be found in Borich, G., The Appraisal of Teaching: Concepts and Process. Reading, MA: Addison-Wesley, 1977 (Chapter 6).

across subtests and pupil sex. In the final stage, they observed subsamples of 17 2nd-grade and 14 3rd-grade teachers the first year and 15 2nd-grade and 13 3rd-grade teachers the second year, using both high- and low-inference coding systems. The primary coding instrument was a multifaceted low-inference measure based upon the Brophy-Good Dyadic Interaction System (1970). Significant process-product relationships were replicated across both years of the study.

2. Stallings, J., and Kaskowitz, D. "Follow-Through Classroom Observation Evaluation, 1972-73." Menlo Park, California: Stanford Research Institute, 1974.

This study was a multi-year effort that examined four 1st-grade and four 3rd-grade classrooms in 26 cities. These classrooms represented five projects in six Follow-Through programs and six projects in a seventh educational program. The goal of Follow-Through was to examine the differential effectiveness of instructional programs based on divergent theories of education and development that had implications for teacher training and evaluation. One 1st-grade and one 3rd-grade non-Follow-Through classroom were selected for comparison at each project site. Using a multifaceted classroom observation instrument, Stallings and Kaskowitz gathered data about classroom environment and teacher process--specifically about seating patterns, the presence and use of equipment and materials, grouping of children, staff, and activities in the classroom, the role of the person who is the focus of classroom interaction, and the type and quality of that interaction. Pupil behavior relating to independence, task persistence, cooperation and questioning was assessed on the same classroom observation system. Reading and math skills were measured on the Metropolitan Achievement Tests, and problem solving and pupil responsibility were assessed on additional paper and pencil measures.

3. Good, T. L., and Grouws, D. A. "Process-Product Relationships in Fourth Grade Mathematics Classes." Columbia, Missouri: College of Education, University of Missouri, 1975.

This study, in many ways similar to the Brophy-Evertson research, examined relationships between teacher process and pupil mathematics achievement in 4th-grade classrooms. Following the method employed by Brophy and Evertson, Good and Grouws selected a subset of 41 teachers from a total sample of 130 whose students had demonstrated gains on the Iowa Tests of Basic Skills for 2 consecutive years. Teacher behavior was measured on two instruments: the low-inference Brophy-Good Dyadic Interaction System, which codes approximately 164 discrete teacher behaviors, and a high-inference scale on which 8 variables (organization, alerting, accountability, classroom climate, thrust of homework, student attention, clarity, and enthusiasm) were rated in a Likert-style format. In analyzing their data, Good and Grouws performed a test of significance between differences in the behavior of the top and bottom nine and the top and bottom three teachers. These relative rankings were established by determining the mean residual pupil gain score for each teacher over 2 consecutive years. Thus, the more and less effective teachers were those whose pupils had the highest positive residual gain scores (top) and highest negative residual gain scores (bottom) over 2 consecutive years on the Iowa Tests of Basic Skills total math subscale.*

4. Soar, R. S. "An Integrative Approach to Classroom Learning." Philadelphia, Temple University, 1966. ERIC Document Reproduction Service (ED 033 749).

Soar, R. S., and Soar, R. N. "An Empirical Analysis of Selected Follow-Through Programs: An Example of a Process Approach to Evaluation." in I. J. Gordon (Ed.), Early Childhood Education. Chicago: National Society for the Study of Education, 1972.

These two citations actually represent a series of four studies. The first

---

*A residual gain score for a particular teacher represents his or her pupils' level of achievement over (positive residual) or under (negative residual) the average gain for all classrooms.

was conducted in four elementary schools, grades 3 through 6. The behavior of 55 teachers was recorded on three observation systems: The Flanders Interaction Analysis System; a second instrument specifically designed to cover areas outside the Flanders exclusive focus on verbal behavior; and a third measure for recording affect--positive and negative, teacher and pupil, verbal and nonverbal. Pupil measures were obtained on the vocabulary, reading, arithmetic concepts, and arithmetic computations subtests of the Iowa Tests of Basic Skills. These were supplemented with a number of personality, attitude, and creativity measures.

The second study was conducted using the Follow-Through data described in the Stallings and Kaskowitz study. Its primary objective was the identification of dimensions of teacher behavior which were related to pupil gain across programs. Eight teachers from each of seven experimental programs were observed, along with two comparison teachers from each program site.

In the third and fourth studies, a 1st-grade sample of 22 classrooms and a 5th-grade sample of 59 classrooms were employed. The observation measures used on these samples, and on the Follow-Through sample above, included a revision of the instrument developed for the first study, in order to code the teacher's classroom management behavior and the pupils' response to that behavior. Another observation instrument employed in the last three studies recorded pupil interaction as comprehensively as the Flanders coded teacher behavior, and a third measured cognitive behavior exclusively.

Like the other studies described, these four focused primarily on the reading and mathematics achievement of pupils and employed residual gain scores corrected for pretest achievement.

5.* McDonald, F. J., Elias, P., Stone, M., Wheeler, P., Lambert, N., Calfee, R., Sandoval, J., Ekstrom, R., and Lockheed, M. "Final Report on Phase II Beginning Teacher Evaluation Study." California Commission on Teacher Preparation and Licensing, Sacramento, California. Princeton: Educational Testing Service, 1975.

This study was the initial phase of a long-term investigation of teacher effectiveness. Pupil performance in reading and mathematics was assessed at two points in time and the teachers' classroom behavior during the intervening period was measured and then related to differential pupil achievement. The California Achievement Test was used to measure pupil cognitive performance, while various other instruments were used to assess pupil attitudes, cognitive style, expectations, and classroom behavior. The performance of 44 2nd-grade and 53 5th-grade teachers with 3 or more years of experience was recorded during reading and mathematics instruction on an observational coding system especially developed for this study. The system included categories for the teacher's introductory remarks, explanations, questions, reactions to pupil behavior, and feedback to pupil learning. Two global rating scales were used to measure teacher feedback, directiveness, management, and classroom control as well as other general behaviors such as motivation, warmth, and honesty. As in the other studies described, teacher data were related to the adjusted posttest achievement scores of pupils in order to identify more effective and less effective teaching behaviors.

Table 1 summarizes the general methodological characteristics of these studies.

---

*The results of Phase III of this study will be available from the California Commission on Teacher Preparation and Licensing, Sacramento, California, in the fall of 1978.

Table 1. Some Contextual Characteristics of Five Major Process-Product Studies

Researchers	*Grades*	*Content*	*Sample size*	*Sample selection method*	*Criterion measures**
Brophy-Evertson	2nd, 3rd	Reading, math	1st year: 17 (2nd); 14 (3rd) 2nd year: 15 (2nd); 13 (3rd)	Self-selected + consistency in producing learning gains over a four-year period	Residualized gain, MAT
Soar	1st, 3rd, 4th, 5th, 6th	Reading, math	Study 1: 55 (3rd-6th) Study 2: 20 (1st) Studies 3 & 4: 22 (1st); 50 (5th)	Self-selected	Residualized gain, ITBS, MRT, MAT
Stallings	1st, 3rd	Reading, math	105 (1st) 98 (3rd)	Self-selected	Raven's, MAT with WRAT as covariable, IAR, SRI observation instrument
Good-Grouws	4th	Math	41	Self-selected + top and bottom on residualized gains	Residualized gain, ITBS
McDonald	2nd, 5th	Reading, math	44 (2nd) 53 (5th)	Self-selected + three years experience	CAT as covariable

* Key to criterion measures: MAT = Metropolitan Achievement Test; ITBS = Iowa Test of Basic Skills; MRT = Metropolitan Readiness Test; WRAT = Wide Range Achievement Test; IAR = Intellectual Achievement Responsibility Scale; SRI = Stanford Research Institute; CAT = California Achievement Test.

The findings from these and similar studies, usually in the form of process-product correlations, comprise the core of competency-based statements. These findings, together with conceptual models of effective teaching, professional experience, and the values and goals of the training institution,provide rationale for the teacher behaviors that are taught. Accordingly, much of the criterion performance measured in evaluation studies is rooted in one or more process-product studies from which training institutions extrapolated competency-based statements. It is important to note that the translation of these findings into competencies is undertaken not by the researcher, but by the teacher trainer. And neither the researcher nor the teacher trainer is likely to specify validated proficiency levels for the behaviors in question. Given the correlational nature of the researcher's methodology, no cause and effect implications can be drawn from his findings. Any proficiency levels that <u>are</u> established probably reflect the values and preferences of specific institutions rather than evidence that the teacher who has achieved them is more likely to engender desirable pupil outcomes than the teacher who has not.

Findings derived from the five studies described above follow. Teacher behaviors are listed within studies in order to convey the number and type of outcomes produced by each research effort--although some findings appear to run across subsets of studies. The behaviors listed represent only those teacher processes, skills, or performances that have exhibited significant relationships to pupil outcomes in mathematics and/or reading. Many other behaviors--often two, three, or even four times the number listed--were studied and found unrelated to pupil achievement. To illustrate the translation of process-product relationships into competency statements, to the left of each teacher variable is listed its most specific implication for preparing a competency statement and establishing a proficiency level. A summary chart at the conclusion of this list indicates consistent and inconsistent findings across studies.

BROPHY-EVERTSON RESULTS

General Findings

Variables	Competency Implication*
1) Classroom Management	(Teacher should have the ability to)** keep pupils actively engaged.
2) Rules	Establish flexible rules sufficient to keep order, and change them when necessary.
3) Punishment	Use mild, non-physical forms of punishment.
4) Role Definition	Take personal responsibility for student learning and have high expectations.
5) Individualization	Match the difficulty of the lesson with the ability of the pupils, and vary the difficulty as necessary.
6) Group Lessons	Call on children systematically rather than randomly. Give students opportunity to practice newly learned concepts and to get feedback.
7) Teacher Feedback	Give credit for partially correct answers. Accept questions in the form they are asked. Give feedback.

*The reader should note that the competency implications are drawn from correlational studies, and thus these variables may be associated with other, yet unidentified, variables that are causal to pupil learning. The experimental manipulation of teacher behavior and random assignment of pupils to teachers are two methodological characteristics required to establish the cause and effect nature of these implications.

While these competencies are derived from interpretations, expressed or implied, made by the researchers in their respective papers, the author takes sole responsibility for their accuracy in embodying the spirit of the researchers' conclusions (as must the teacher-trainer when following this same process).

**This phrase is implied throughout the remainder of the list. Teacher behaviors specific to a particular type of student and/or dependent variable are indicated by variable headings or the wording of the competency implication.

Variables	Competency Implication
8) Student Initiation	Encourage students to ask questions.
Findings for Low-SES Pupils	
9) Teacher Affect	Be warm and encouraging, let students know that help is available.
10) Student Responses	Elicit a response from the student each time a question is asked, before moving to next student or question.
11) Over Teaching/Over Learning	Present material in small chunks, at a slow pace, with opportunity for practice.
12) Classroom Interaction	Stress factual knowledge. Monitor student progress. Minimize interruptions by maintaining smooth flow from one activity to another. Help student who needs help immediately.
13) Individualization	Supplement standard curriculum with specialized material to meet the needs of individual students.
Findings for High-SES Pupils	
14) Praise and Criticism	Correct poor answers when student fails to perform.
15) Individualization	Ask difficult questions. Follow prescribed curriculum. Assign homework.
16) Classroom Management	Be flexible. Let students initiate teacher-student interaction. Encourage students to reason out correct answer.
17) Verbal Activities	Engage students in verbal questions and answers.

Variables	Competency Implication
STALLINGS-KASKOWITZ RESULTS	
1) Length of School Day	Maximize instructional time.
2) Systematic Instructional Patterns	Use this instructional model: (a) provide information, (b) ask questions about the information, (c) allow child to respond, (d) give feedback, and (e) guide pupil to correct response.
3) Discussion	Encourage class discussion of material during mathematics instruction.
4) Praise	Encourage and praise pupil with low entering ability during mathematics instruction.
5) Textbooks and Programmed Workbooks	Use textbooks and programmed workbooks during mathematics instruction and foster task persistence.
6) Flexible Classroom	Use wide variety of materials and audio-visual aids. Be flexible in allowing pupils to select groups and seats. Encourage pupils to take responsibility for their success.
7) Exploratory Materials	Use variety of exploratory materials to foster pupil cooperation.
8) Question Asking	Respond to child questions and engage in conversations with the child.
GOOD-GROUWS RESULTS*	
1) Whole Class Instruction	Maximize time class is taught as a single unit.

*The results reported are those that are consistent across separate analyses of data obtained for the top and bottom nine teachers and the top and bottom three teachers.

Variables	Competency Implication
2) Classroom Climate	Reduce classroom tension and anxiety. Engender relaxed, non-evaluative classroom atmosphere.
3) Accountability	Establish pupil standards and expect specific pupil accomplishments.
4) Feedback	Give students as much information as needed, especially through process feedback.
5) Questioning	Ask unambiguous questions that pupil can answer completely or not at all. If pupil cannot answer question, supply information about how it might be answered.
6) Praise	Limit praise, especially when performance is poor and expectations low.
7) Teacher Initiated Contact	Avoid approaching specific pupils for the purpose of criticizing or blaming.
8) Pupil Initiated Contact	Encourage pupils to approach teacher individually on work-related matters.
9) Classroom Discipline and Management	Maintain classroom free of major behavioral disorders.
SOAR-SOAR RESULTS	
1) Direction and Control of Learning	Provide moderate direction and control of pupil learning, avoiding extremes.
2) Structuring	Vary amount of structure; reduce structure for more complex content (high cognitive objectives); and increase structure for more elementary content (low cognitive objectives).

Variables	Competency Implication
3) Teacher-Pupil Interaction	Vary level of teacher-pupil interaction, depending on pupil's ability to cope successfully with the activity at hand. Match level of interaction with difficulty of activity, reducing teacher-pupil interaction when pupils are not coping successfully.
4) Teacher Affect	Vary level of affect: Increase positive affect for low-SES pupils, reduce positive affect for high-SES pupils.
MCDONALD ET AL. RESULTS	
1) Instruction Time	Maximize direct instruction time during reading by using group procedures, while maintaining a high level of interaction with individual pupils (2nd).* Increase individual monitoring and reduce group work during mathematics instruction (2nd).
2) Instructional Content	Maximize coverage of instructional content per unit of time during mathematics instruction (2nd, 5th).
3) Instructional Activity	Devote considerable time to discussing, explaining, questioning, and stimulating cognitive processes during reading instruction (5th). Minimize the use of instructional materials that may complicate management of instruction (5th).
4) Instructional Organization	Maximize group work during mathematics instruction (5th). Minimize independent work during mathematics instruction, especially that which may interfere with on-task behavior (5th).

*The codes (2nd) and (5th) refer to the grade levels at which these implications are most applicable.

Variables	Competency Implication
5) Interactive Techniques	Employ specific cues and questions that require the student to attempt a response during reading instruction (2nd). Employ thought-provoking questions during reading instruction (5th).

General Findings and Policy Interpretations

The above studies--and, in fact, most process-product investigations from which competency statements are derived--share the following characteristics:

1. They are confined to early and mid elementary grades and primarily to reading and mathematics instruction.

2. They focus on pupil outcome as measured by nationally standardized tests of pupil achievement.

3. They emphasize teacher behaviors measurable on low-inference classroom observation systems.

4. They have produced qualified findings within SES level, grade, and subject matter.

5. They have been conducted with experienced teachers.

In addition, these studies share many (but not all) findings. Both congruent and incongruent results are summarized below and in Table 2.

1. Teacher Questioning. The value of a systematic, patterned questioning strategy that focuses on individual pupil needs and understanding was confirmed in both studies that investigated this variable.

2. Whole Class Instruction. The value of teaching the class as a unit was confirmed in two out of three studies.

3. Instructional Materials. The value of using specialized materials, including textbooks and workbooks, was confirmed in two out of three studies.

4. Praise. The value of praise was unclear, though it appeared to be related to pupil SES, with lower-SES pupils profiting more from this teacher behavior than higher-SES pupils.

5. Flexibility. The value of flexible rules was confirmed in both studies that investigated this variable.

6. Control and Structuring. The value of controlling pupil responses and structuring pupil behavior was confounded by pupil SES and the cognitive objectives of the teacher. Lower-SES pupils benefit from tighter control, and higher cognitive objectives are more likely to be achieved in a less structured situation.

7. Interaction. The value of teacher-pupil interaction may depend on the situation and the kind of interaction. It appears to have a positive effect during group lessons and a negative effect most other times.

8. Teacher Affect. The value of high teacher affect with low-SES pupils and low teacher affect with high-SES pupils was confirmed in both studies that investigated this variable.

9. Pupil Engagement. The value of engaging pupils in on-task behavior (and keeping them engaged) was confirmed in two out of three studies, and may have been situation-specific in the third.

10. Student-Initiated Questions. The value of student-initiated questions was confirmed in both studies that investigated this variable.

11. Clarity. The value of teacher clarity was unapparent, and probably content- and situation-specific.

12. Attention Getting. The value of getting and keeping pupil attention was confirmed in both studies that investigated this variable.

13. Feedback. The value of feedback seems to have been related to the aspect of pupil performance (substance or form) to which the teacher was responding. Feedback on substance had positive impact on pupil achievement, while the effect of process feedback depended on its context and type.

These findings have policy implications for teacher educators and evaluators and, of course, must be interpreted in light of societal and professional values. To keep them within the realm of general (albeit not universal) consensus, the following guidelines are offered.

1. The range of competencies to be exhibited by teachers should include both broad contextual behaviors (e.g., ability to select appropriate strategies according to pupil SES, and level of cognitive objective) as well as basic skills and behaviors (e.g., probing, questioning, reinforcing) that are likely to generalize across contexts and populations.

2. Any list of competencies should include both performance behaviors, observable in the classroom, and their prerequisite knowledge behaviors.

3. Systematic inservice training of teachers should go hand-in-hand with the assessment of teacher competency. Since there is no consensus about the best method of

training (e.g., self-paced training modules, workshops, graduate school, etc.) a variety of alternatives should be explored and evaluated, and none fostered to the exclusion of all others.

4. The primary purpose of competency-based assessment should be to stimulate the professional growth and development of the individual teacher. Thus, any evaluation related to individual teacher performance should include the means by which to remediate weaknesses identified.

The remaining portion of this paper will identify three evaluation models with which the above competency implications can be used.

Table 2. Selected congruent and discrepant findings for 5 research studies.

Brophy-Evertson	*Soar*	*Stallings*	*Good-Grouws*	*McDonald*
Teacher responds to each question + L*		Provides information/ asks question (systematic instructional pattern) +		
Making sure student understands + L*				
		Use of small groups +	Teaching whole class +	Teaching whole class +
Specialized materials + L		Use of textbooks and workbooks +		Variety of instructional materials −
Praise after student answers opinion questions + L		Praise**	Praise −	
Student initiated praise − L				
Flexibility of rules +		Flexible classrooms +		
Controlling student responses + L, − H	Direction and control of learning ∩ ***			Time organizing instructional activity −
Teacher structuring and feedback − L	Unobtrusive structuring behavior − L, + H			
Interacting with individuals during group lessons +	Teacher-pupil interaction at high cognitive level −		Teacher afforded contact with students −	
Teacher affect + L, 0 H	Teacher affect + L, − H			
Keeping students actively engaged +				Maintaining task involvement −
			Time teaching whole class +	Content covered +
Student initiated questions +			Student initiated interaction +	
Clarity 0			Clarity +	
Getting groups' attention +			Alerting behavior +	
Giving student correct answer +			Process feedback −	
Responding to substance rather than form +				
Failure to give feedback −				

Note: + indicates positive relationship to pupil achievement, − indicates negative relationship, 0 indicates no relationship.

* L indicates finding for low-SES pupils only, H indicates finding for high-SES pupils only.

** The effect of praise on achievement in math in first grade was variable: in classrooms where children had relatively low entering ability, pupils profited more from a high rate of praise than they did in classrooms where students had higher entering ability.

*** Soar's inverted U; indicating a curvilinear relationship between direction and control of learning and pupil achievement.

Evaluation Models: Process and Strategy

Competencies derived from process-product studies have contributed in large measure to the content of inservice evaluations in the U.S. However, because of their limited intent, i.e., to investigate process-product relationships, these studies have been less useful in suggesting methodologies by which to conduct inservice evaluations. Hence, three models are presented in this section in order to convey the variety of methodologies commonly employed in teacher evaluations. These models are neither the only nor necessarily the best available. Indeed, the "best" evaluation methodology is dictated by context and dependent upon resources at hand, time and commitment of those conducting the study, requirements and policies shaping the evaluation, and, of course, the objectives of the training institution.

The models are presented in an order that reflects the time and expense generally required for their implementation, with the least costly and time-consuming appearing first. For purposes of this paper, the three models are titled Needs Assessment, Relative Gain, and Process-Product. They have no commonly accepted titles--probably because they are generic, or pure, models, rarely implemented in full or in the precise form illustrated here. Yet they contain the basic ingredients from which most, if not all, other evaluation designs are constructed.

The Needs Assessment Model for Evaluating Inservice Education and Training

A training need is defined as the discrepancy between an educational goal and actual teacher performance in relation to this goal. Thus, needs are expressed in terms of the teacher, not the teacher trainer or administrator. Moreover, they are expressed not as teaching resources (teaching aids or

or supplementary materials), but as behaviors or skills to be attained. In order to improve both the individual teacher's classroom performance and the training program itself, the trainer must specify for each behavior the level of proficiency at which the teacher is considered "competent." Hence, these behaviors and skills are referred to as competencies.

The needs assessment model begins with a set of competencies. These may be competency implications derived from process-product studies or unvalidated behaviors and skills drawn from the professional experience of teacher trainers. Thus, while both types of competency focus on teacher performance, their origins differ: the latter, which often sound more like goals than competencies, stem from the professional judgment of trainers and developers, and the former from empirical findings of teacher effectiveness research employing the criterion of pupil achievement.

In the case of goal-based competencies, considerable time may be required to formulate a comprehensive list of goals with which all interested parties (trainers, developers, administrators, etc.) can agree. Selecting goals can be a complex process since policies and priorities at both institutional and state levels must be considered together with regional needs and the implicit and explicit values of teacher trainers. This process, while logically preceding development of the training program, is often prompted by a perusal of the materials and activities already developed for training and presumably embodying training goals.

The needs assessment model serves the evaluator by identifying discrepancies between the competencies the teacher should possess and those the teacher believes he or she possesses. In this respect, the model incorporates the self-report characteristic of follow-up studies--a fact which may account, in part, for its popularity among cost-conscious training

institutions. The evaluator uses these discrepancies to ascertain the effectiveness of the training program in teaching each of the competencies, and, specifically, to pinpoint components of training that are not engendering the intended behavioral outcomes.

Inservice teachers are polled about their ability to perform the competencies taught as soon as a reasonable amount of time has elapsed following training. After short-term training the newly acquired skills of the teacher may be tested immediately. After degree or certification programs, assessment may occur anytime during or throughout the next semester or year, the assumption being that the more complex the training objective, the more time needed for evaluating the behaviors taught. When extensive training has been provided, the training institution may observe the teacher in situ and record his or her competence in performing selected behaviors. While classroom observation may raise the cost of the evaluation, data obtained in this manner supplement and corroborate self-report information.

Following are the steps involved in implementing the needs assessment model.

1. List competencies. Competency statements are derived either from process-product studies or from the intents and objectives of teacher trainers, or both. Inservice teachers may assist in this task by supplying "competencies" in an open-ended fashion after being given a definition of "competency" and shown examples written in an acceptable format. Competencies on master lists are sorted for the purpose of selecting topics for training. Afterwards, the newly developed training program and/or materials are examined to insure that selected competencies are actually translated into program activities and materials. This list of training topics is used in constructing the survey instrument.

2. Survey inservice teachers. All or a sample of inservice teachers who have completed training are asked to rate (a) the relevance of each competency to their current job function (or perceived

future job function) and (b) their current attainment of each competency. A typical questionnaire might take the following form:

Competency	Perceived Relevance				Perceived Level of Attainment			
	Low			High	Low			High
1.	1	2	3	4	1	2	3	4
2.	1	2	3	4	1	2	3	4
3.	1	2	3	4	1	2	3	4
4.	1	2	3	4	1	2	3	4

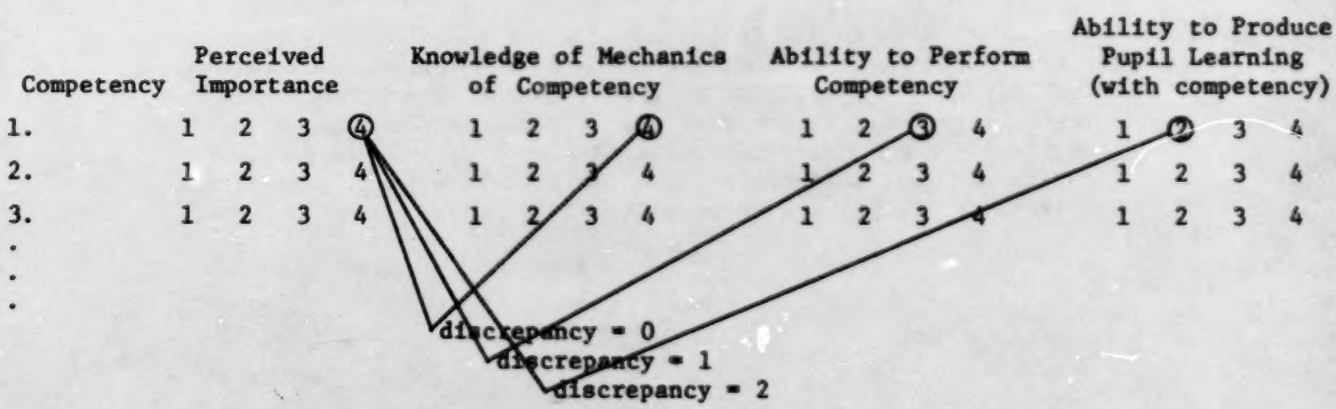
A more exact (but less common) way to rate competency attainment is to divide each competency statement into "knowledge" competence, "performance" competence, and "consequence" competence. These terms might be defined on the questionnaire in the following manner.

Knowledge competence: Ability to accurately recall, paraphrase, or summarize the procedural mechanics of the behavior on a paper and pencil test.

Performance competence: Ability to accurately execute the behavior in a real or simulated environment in the presence of an observer.

Consequence competence: Ability to elicit learning from pupils (as recorded on tests of affective and/or cognitive achievement) by using the behavior.

These distinctions require the teacher to make finer judgments in rating each competency and in turn permit a more refined evaluation of the training program. Questionnaires incorporating these competency dimensions might take the following form.



Each competency then yields for each respondent three discrepancy scores which indicate the effectiveness of the training program in producing (a) trainee knowledge, (b) trainee performance, and (c) pupil consequences. Using these discrepancy scores as a guide, components of the training program can be revised to produce greater "knowledge,"

"performance," and/or "pupil impact." The three discrepancy scores above indicate that field experiences for this competency (performance and consequence) need improvement but classroom instruction (knowledge) is adequate.

3. Rank Competencies. Competencies are then ranked according to ratings obtained on the above questionnaire. For each competency the difference between perceived importance and perceived level of attainment is calculated across the three dimensions, knowledge, performance, and consequence. These differences are ordered according to magnitude or relative weight (based on average perceived importance determined over all respondents x average discrepancy). If the average perceived importance of competency 1 were 2.5, the resulting knowledge discrepancy would be 0, the resulting performance discrepancy would be 2.5, and the resulting consequence discrepancy would be 5.0. Other competencies deemed either more or less important than this competency would have their discrepancies adjusted accordingly. This weighted ordering takes into account that a small discrepancy on one competency may be of greater perceived importance than a large discrepancy on another competency. Those discrepancies with the greatest positive rank difference should have the highest priority in regard to revising the training program.

4. Compare High Priority Competencies with the Content of the Training Program. High priority competencies determined from the above analysis are compared to the instructional experiences, components, and materials that receive high priority in the training curriculum. The evaluator might look at instructional time devoted to the competency, clarity of instruction, adequacy of training materials, and number of minutes or hours allotted for practicing the competency in order to determine the emphasis that the training program places on a given competency. When a competency is highly valued but poorly performed, the problem may stem from insufficient rather than ineffective training.

5. Revise Program or Revise Competency. Where possible, the emphasis of the training program is modified to match high priority competencies. Or, if altering the training program to stress a particular high priority competence is not cost effective, other training resources (e.g., self-paced modules, programmed texts) or sources (e.g., agencies and institutions at which the inservice teacher may obtain the needed training) are recommended.

The needs assessment model is commonly extended and adapted to meet the user's particular needs. For example, the needs assessment instrument is sometimes used in conjunction with a similar survey completed by supervisors or administrators in order to corroborate the subjective responses of inservice teachers. An evaluation of training, for instance, might be based on the mean discrepancy across teachers and supervisors, thereby taking into account a second and presumably more objective group of respondents. Or, competencies for which the reported level of attainment differs dramatically from supervisor to teacher may be withheld from analysis pending clarification by data from other sources, such as in situ observation.

Evaluations using the needs assessment model generally have multiple purposes also. These purposes derive directly from the nature of discrepancy data, which can be used with equal effectiveness for either summative or initial judgments about training. Data revealing the perceived importance of the competencies studied, for example, can serve both as a check on the relevance of training and a guide to the development of additional training. The versatility of these data make the needs assessment model less evaluative, less restrictive, and more developmental than most other approaches to the evaluation of training.

It is important to emphasize that a needs assessment is essentially a self-evaluative procedure which relies heavily on judgments of teachers about their own performance. This characteristic, which makes the model very appealing to the teacher, is considered a weakness by the evaluator. Thus, efforts to strengthen the needs assessment model often include supervisor-administrator ratings or limited follow-up visitation to enhance the credibility of self-reports and to obtain additional vantage

points from which to judge discrepancies between program intents and the post-training performance of teachers.

While this model receives almost universal support, its strongest proponents are teachers themselves and the associations that represent them. This advocacy is based on the assumption that the performer (teacher) can best judge his or her own performance and, when explicitly asked to do so, can make an objective judgment. This assumption is more tenable, however, when the purpose of data collection is the evaluation of prior training, not assessment of individual teachers. Proponents of program accountability argue that this distinction is always blurred regardless of efforts to the contrary and are quick to point out evidence that response bias differs systematically depending on (1) whether one is rating individual or group performance and (2) whether one is a member of the group one is rating. Accountability proponents have also questioned the ability of the teacher to accurately judge the complex behaviors that comprise competencies. Such judgments, it has been suggested, may require special training due to the vague and general terminology often used in competency statements. Thus, compromise usually prevails in the implementation of the needs assessment model; the methodology is extended wherever possible to other samples that are assumed capable of corroborating the teacher's judgment.

An important practical characteristic of the needs assessment model is the ease with which it can be implemented by nontechnical personnel. It is sufficiently direct that data analysis and instrument construction are no more complex in a needs assessment than they are in a follow-up survey. Consequently, it is the model most often implemented by teacher trainers and program developers who want immediate feedback on the effectiveness of

program experiences and materials. It is even more popular with those who have limited resources, since it can be conducted along with the ongoing training activities of a school, college, or department of education with little or no additional staff, funds, or equipment. Professional evaluators generally are not among those implementing the needs assessment model. While the presence of an evaluator might facilitate implementation of the needs assessment, it probably would not affect the quality of the data, which is to a great extent fixed by the nature of the model. The evaluator's presence is more appropriate when a federal or state agency is funding the evaluation, in which case it is likely to be more summative than formative, emphasizing the effectiveness of previously developed training rather than the creation or modification of training.

Finally, we must note the definition of evaluation implied by the needs assessment model: the process of determining the congruence between what "should be" and "what is," i.e., between what the teacher should be able to do and what the teacher can do. The evaluation is complete when the training program has objectively determined the discrepancy between these two poles. This definition, then, calls for the development of goals and objectives (in the form of competencies) and an assessment to measure the extent to which these goals and objectives have been met. Generally, this is accomplished by obtaining self-report data from trained teachers about the value and attainment of each training objective.

The Relative Gain Model for Evaluating Inservice Education and Training

Impetus for the relative gain model has come from the belief among U.S. taxpayers that teachers, principals, and educational programs should

be accountable to their constituencies. This belief has prompted a number of states to pass laws making evaluation and accountability procedures mandatory at the school and school district level. In response to this mandate, procedures have been developed to compare student performance in different classrooms in an effort to establish minimum standards of pupil growth for which all teachers can be held accountable.

The primary assumption of the relative gain model is that a truly effective educational system emphasizes objective assessment of teachers. Proponents of educational accountability often find the needs assessment model too susceptible to individual bias to provide accurate data upon which to base accountability decisions. They feel a more valid index of teacher effectiveness is pupil achievement. If a teacher is doing his or her job well, that teacher's pupils should exhibit more learning than those taught by a teacher who is not doing his or her job well. This assumption, of course, is reasonable only if (1) those factors over which the teacher has no influence but which affect pupil performance are controlled, and (2) the phrase "doing his or her job well" is translated into meaningful units of pupil achievement. To resolve the first problem, pupil achievement scores must be adjusted to account for differences among pupils prior to instruction. To resolve the second problem, traditionally the more difficult of the two, pupil tests that are sensitive to the competencies stressed in the training program are needed.

By focusing on pupil performance, the relative gain model measures behavior at least one step removed from the training program. The effect of training must register not only on teacher measures, but also on tests of pupil performance. The relative gain model rests on the assumption that teacher competencies can be translated directly into pupil competencies

and that potentially confounding variables can be statistically controlled to an extent sufficient to allow the effects of teacher training to filter down to the pupil.

Critics of the relative gain model complain that the path from teacher trainer to pupil is long and full of potential obstacles. Yet there are two procedures that can be used to increase the likelihood of obtaining valid measures of relative gain. The first involves the use of criterion-referenced rather than norm referenced tests. Nationally normed tests provide only a single score on very general objectives; their content is sometimes irrelevant to the specific objectives of a particular training program. When a nationally normed test is used as the measure of pupil gain, a teacher who exhibited the required competencies at the conclusion of training may find that the expected pupil effects are not measured by the only index of program effectiveness used. For this reason, criterion-referenced tests of pupil performance are recommended for the relative gain model. These tests are designed to measure only the pupil outcomes that are related to content areas for which training has been provided. They are usually constructed by training program personnel, who can control the time of their administration to suit the purposes of the program. Norm-referenced tests, on the other hand, are usually administered only once a year, on a prespecified date.

A second procedure that helps to align training and testing is the performance test. Analagous to a job sample, the performance test is designed to assess, as efficiently as possible, specific, discrete units of a training program. Students are assigned randomly to teachers. The teachers then give a mini-lesson that is designed to

engender specific pupil outcomes related to a particular competency for which training has been provided. Pupils are tested immediately upon completion of the lesson, which is usually given at a critical juncture in or just after training. Performance tests, like criterion-referenced tests, are constructed by the training program, but unlike the latter, are usually administered in a controlled setting prior to the teachers' return to the classroom.

Another procedure that strengthens the link between teacher training and pupil outcome is the clustering of competencies. Here, competencies are grouped according to their expected effects upon pupil behavior. Criterion-referenced or performance tests are then constructed to measure the joint or interactive effect of a competency cluster, thereby permitting a larger chunk of training to be evaluated at one time. While the attainment of a single competency may engender little, if any, change in pupil growth, the attainment of several may substantially heighten the effect that can be logically expected and measured in pupils. Both competency clustering and criterion-referenced testing have become accepted elements of the relative gain model.

Following are the steps used in implementing the relative gain model.

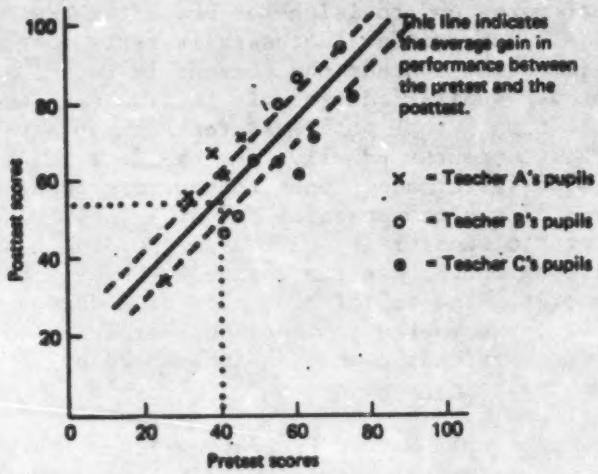
1. List competencies. As in the case of a needs assessment, implementation begins with identification of competencies to be measured. These can be derived directly from a study of the materials and transactions within the training program or from a list of program objectives. The latter source is considered less reliable because the program may not be teaching all that it is said to teach. But the former source may require considerably more time and expense, particularly if it involves direct observation and study of program materials. Competencies derived from process-product studies, especially those that have been shown to relate positively and significantly to pupil outcomes, are usually given highest priority in the relative gain model.

2. Construct Measures of Pupil Performance. The importance of Step 1 to the success of the relative gain model cannot be overemphasized. Because the model makes no provision for measuring teacher behavior, the validity of all outcome data necessarily rests on the assumed relationship between the content the teacher is taught and the behaviors his or her pupils acquire. It is from the competency lists compiled in Step 1 that post measures of pupil performance are derived. These measures usually take the form of criterion-referenced tests or performance tests that assess pupil behaviors logically related to the competencies taught. Their construction requires considerable sensitivity on the part of the test developer in deciding what pupil outcomes can logically be expected from what teacher competencies. The validity of this procedure depends both on the clarity of the competencies and their capacity to engender specific pupil outcomes that cannot be influenced by factors unrelated to the training program.

3. Collect Pre and Post Pupil Data. Pretesting is an indispensible component of the relative gain model. It is the evaluator's hedge against the criticism that pupil performance is, in part, determined by factors out of the teacher's control. The pretest performance of pupils in the form of achievement and/or aptitude scores represents the covariable data with which posttest achievement must be statistically adjusted if the relative gain model is to measure pupil effects that are truly teacher specific. While pretest achievement data obtained on an alternate form criterion-referenced test are most frequently used, the model is sufficiently general to accommodate any number of covariables, including pupil attitudes, interests, and previous experiences, that might otherwise confound the measurement and interpretation of teacher effects. To the extent that the evaluation ignores pupil variables that correlate with posttest pupil achievement, the accuracy and appropriateness of the relative gain model can be called into question. Pre and post achievement tests are used to assess each pupil's performance on all or a large sample of the relevant competencies within a given grade or age level. Generally, different sets of items are used for pre- and posttests to minimize the reactive effects of testing on pupil learning.

4. Plot Teacher Effects. Statistical analysis for the relative gain model begins with the construction of a prediction or regression equation between pre- and posttest pupil performance. This equation may utilize any number of covariables including both achievement and aptitude variables. For illustrative purposes, however, this discussion will be limited to the simplest case, that of pretest pupil achievement regressed on posttest achievement. First, pretest and posttest scores are plotted for each pupil within a given grade or age level. The pupils' teacher is also identified in this process. The scores of five pupils in each of three classes are plotted below to illustrate the procedure.

In this figure* two of the teachers participated in a training program while the third did not.



Next, a regression line representing the average or typical relationship between pretest and posttest is drawn among these points on the plot. A standard linear prediction equation is used for this purpose. The evaluator uses this equation to determine the typical posttest score for each pupil's pretest score. This is indicated by the dotted line (....) for a given pretest value. A typical equation might take the following form:

> Pupil's score on CRT at end of lesson = 1.4 (CRT score at beginning lesson) + 1.75

If, for example, a pupil's score at the beginning of the lesson was 26 points, then the expected score at the end of the lesson would be 38.15, i.e., 1.4 (26) + 1.75 = 38.15. The regression equation describes the best relationship between input and outcome measures. It is positioned through the data points in such a way that the evaluator will make the fewest errors in predicting posttest scores from pretest scores.

This general approach can be expanded to accommodate more than a single index of pupil entry behavior. When this is done, the procedure is called "multiple regression," and the equation takes the form:

> Pupil's score on CRT at end of lesson = b_1 [score on first entry measure] + b_2 [score on second entry measure] + Constant

---

*Adapted from Klein, S., and Alkin, M. "Evaluating Teachers for Outcome Accountability," in Borich, G. The Appraisal of Teaching: Concepts and Process. Reading, Mass.: Addison-Wesley, 1977.

The "b's" in the equation represent weights used to adjust the posttest measure according to the pupil's performance on the pretest and its relationship to the posttest. The number of pre-measures that may be employed is determined by the combination of measures that are correlated with the posttest but not with each other. Generally, only a few pre-measures will meet this criterion. That is, only a few will contain unique information not supplied by other variables already entered in the equation.

5. Construct Confidence Band. When the actual outcome score for a given pupil is greater than the score predicted, performance is said to be "above" expectancy. Similarly, if the posttest score is lower than predicted, the pupil is said to be "below" expectancy. Above and below expectancy are relative determinations, since the standard for "above" and "below" is derived from a comparison of each pupil's actual pre- to posttest improvement with the average improvement of all pupils. Hence, the name, "relative gain model." Just how much an actual score must deviate from the expected score in order for it to be classified "above" or "below" is determined by constructing a confidence interval around expected scores. This band is illustrated above by the broken line on either side of the regression line. This band is a function of the standard error of estimate of the regression equation, and its width can vary. The usual procedure is to use a band that is wide enough to accommodate approximately 2/3 of the pupils.

6. Construct Summary Table. Data from the relative gain model are reported in a table which takes the following form*:

	Teacher A	*Teacher B*	*Teacher C*
Average pretest score	35	50	65
Average posttest score	57	65	73
Expected posttest score	50	65	80
Difference between expected and actual posttest scores	7	0	7
Percent of pupils who are:			
Above expectancy	60%	40%	20%
At expectancy	0	10%	0
Below expectancy	20%	40%	80%

*Adapted from Klein, S., and Alkin, M. "Evaluating Teachers for Outcome Accountability," in Borich, G. The Appraisal of Teaching: Concepts and Process. Reading, Mass.: Addison-Wesley, 1977.

Data are summarized by teacher and are thus averaged over all pupils for each class. Rows in the table indicate average pre, post, and expected post data and, most important, the percentage of pupils who scored above, below, or at expectancy. It is this latter set of data that is used to evaluate the training program. Because the criterion-referenced instrument was specifically constructed to match the objectives of the training program, most of the pupils of those teachers who received instruction are expected to fall at or above expectancy, and most of the pupils of those teachers who did not receive instruction are expected to fall below expectancy. There will be exceptions in either direction, but the average performance for each class should clearly show the above trend if the training program was effective in teaching the intended competencies and if the teachers employed these competencies in the classroom. If this trend were reversed or if no discernable trend were apparent, the effectiveness of the training program, the appropriateness of the posttest measure, and the selection of entry level measures could be called into question. There is no substitute in the relative gain model for the evaluator's diligence in ruling out the two latter possibilities.

The major advantages of the relative gain model, according to its proponents, are its focus on the ultimate target of the educational system, pupil performance, and its consideration of all pupil entry behaviors which might confound a test of training effectiveness. The model is limited, however, by its need for pupil performance tests in two alternate forms, its inability to compare teachers except on a grade-by-grade or subject-by-subject basis, and its requirement that classrooms evaluated contain approximately 20 or more students in order to arrive at stable estimates of expected pupil performance. Thus, the relative gain model, while serving as a stringent accountability measure, usually requires considerably more time and expertise to implement than the needs assessment model. This latter point raises two questions: how often should the relative gain model be implemented, and on what unit of analysis--classroom, school, or school district--should it be based?

In response to the first question, the relative gain model should be used continuously, or as often as possible, to check the effectiveness of a training program. As pupils change from year to year along dimensions related to their achievement, so will the pitch or slope of the regression line on which the relative gain model is based. And, as this slope changes, the proportion of pupils falling above, at, and below expectancy will also change. Each set of teachers exposed to the training program return to different schools, different pupils, and perhaps vastly different school districts. These variables may be sufficiently related to pre and post measures of pupil achievement to show the program in a different light. If program effectiveness varies considerably across groups of trainees, the characteristics on which the groups vary, e.g., pupil SES, school district, subject area, etc., should be included in subsequent evaluations as pre-measures in the regression equation. This flexibility is an appealing characteristic of the relative gain model; it can accommodate new information about pupils, teachers, or schools which might otherwise confound an evaluation study.

In response to the second question, the relative gain model can be based on all three units of analyses--classroom, school, school district--provided a sufficient number of classrooms, schools, or school districts are included in the study. If training is limited to a single school, with only a few teachers participating, the relative gain model requires that classrooms contain at least 20 pupils each and results be generalized only to teachers within that specific grade in that school. When training is presented across schools, but again within grades, classroom averages are employed. For dissemination and training projects within school districts, pupil performance

can be aggregated by school, and for national projects, where training materials are widely disseminated, it can be aggregated by district. It is important in larger studies that the potentially confounding characteristics of schools and school districts be quantified and added to the regression equation. Finally, separate equations should be constructed for each posttest measure, e.g., affective and cognitive, for each grade level at which the material or program is tested.

The relative gain model reflects the spirit of the accountability movement to a greater extent than the needs assessment model, but to a lesser extent than the process-product model. For this reason, the relative gain model is often considered an intermediate or "middle of the road" procedure for school districts, colleges, and departments of education that wish to employ an approach presumably more objective and quantitative than the needs assessment model, but less stringent than the process-product model. Thus, the relative gain model is best viewed as the midpoint on a scale of objectivity and quantification, falling between the needs assessment and the process-product model.

Depending on the size and composition of the training program, an evaluation may use all three models in the course of a multiyear cycle, or may become fixed for cost or political reasons at the needs assessment stage. Typically, the relative gain model leads to experimentation with teacher observation and its connection to pupil outcome, and, hence, to the process-product model. Thus, the relative gain model is often a transitional step for those institutions that have employed self-report and survey evaluations in the past and wish to experiment with more pupil-oriented procedures.

Few institutions, it seems, remain at the relative gain stage. In some cases, the link between teacher training and student gain is considered far

too tenuous and too confounded by other variables to warrant further study of pupil outcomes. In other evaluations, particularly those that have revealed considerable variance in pupil outcome between trained and untrained teachers, the process-product link is enticing. Thus, the relative gain model often represents a point of decision in the evaluation of training. For this reason, it is likely to be chosen by institutions that can afford the risk of a negative evaluation (e.g., older, more established training programs that have experienced success with some variation of the needs assessment model). For newer training programs whose funding might be threatened by negative findings, the risk may be too high.

The relative gain model may also be used more frequently by the school district conducting inservice training than by the school, college, or department of education since there is greater pressure for accountability at the community level. The autonomy of institutions of higher learning often insulates them from the accountability pressures felt by local school districts. Furthermore, the business of university-based training has traditionally been teacher, not pupil, education.* Thus, schools, colleges, and departments of education have lagged behind school districts in employing the relative gain model. While this "lag" may be diminishing, institutions of higher learning continue to view follow-up studies as an acceptable method of determining the worth of their programs.

More often than not, the relative gain model is threatening to the teacher. This feeling is exacerbated by the fact that the teacher plays no role in study design, implementation, or data collection. In a needs assessment,

---

*Even the highly publicized competency-based movement has dealt primarily with the competencies of trainees (not pupils).

teacher opinion is elicited at several points, but the relative gain model bypasses the teacher to focus directly on pupil performance. Interestingly, when questioned about the source of threat posed by the relative gain model, teachers are quick to cite the variables they cannot control but which they know will affect pupil performance. Here, the inservice teacher can help the evaluator by identifying covariables that should be included in the regression analysis and, hence, "controlled," to insure that pupil outcomes reflect only activity of the teacher.

The threat posed by the relative gain model may be more imagined than real since classroom data are pooled in order to make decisions about the training program. Although the teacher must be identified in this process, it is not the individual teacher on whom decisions are made, but instead the entire group of teachers who have received training.

While the needs assessment model can be implemented entirely by the training staff, the relative gain model is more complex. Its implementation requires at least one staff member with a working knowledge of multiple regression and a computer program to carry out the regression analysis. The process should not be viewed as too complex for educational personnel who ordinarily deal with the data analysis. Many of these individuals are already familiar with the technique, and those who are not can easily learn it. Training staff themselves can and should become proficient at both computing and interpreting the results of a regression analysis. Some initial expertise may be required to implement the relative gain model, but this expertise can usually be found among personnel working in training programs or in school districts.

The characteristic which perhaps most distinguishes the needs assessment and relative gain models is the definition of evaluation implied by each.

In the needs assessment model, evaluation is the assessment of "what is" and "what should be," with the teacher and/or supervisor providing data from which to determine the congruence between these two measures. In the relative gain model, evaluation is the assessment of normative improvement in pupil performance. By "normative improvement" is meant relative pupil gain. Whether a particular pupil's score (or a class mean) is "above" or "below" expectancy depends not on the absolute value of that score but on its value relative to the average improvement of all pupils (classes). By definition, pupils' scores fall both above and below the regression line. Hence, there is always a forced distribution or relative ranking of scores. This ranking will occur regardless of how high or low the pupils score on the pretest or how little variability exists among scores. The relative gain model does not test the effectiveness of training, but rather its ability to discriminate teachers who participated in training from those who did not. Hence, the data reflect gains made by pupils of one group of teachers relative to those made by pupils of another group of teachers.

Finally, by focusing on pupil change, the relative gain model incorporates Scriven's definition of "pay off" evaluation. Because program effectiveness is determined by the ultimate "consumers" or "products" of teacher training, the relative gain model goes considerably further than the needs assessment in measuring the "payoff" expected from training. The fact that this model measures pupil learning gains, taking into account entering level of pupil performance, makes it attractive to teacher trainers, program developers, and sponsoring agencies who wish to determine the extent of their impact on pupils.

The Process-Product Model For Evaluating Inservice Education and Training

Of our three models for the evaluation of training, the process-product approach is the most complex. It is a hybrid, reflecting in some respects the domain of the evaluator and in other respects the domain of the researcher. Because it links these two domains, the process-product model has a unique capacity to serve both the teacher trainer and teacher researcher. This versatility does not come without a price, however--that price being the requirement that both the process behavior of the teacher and the performance behavior of pupils be measured.

Whether the process-product model serves the teacher trainer or the teacher researcher depends on the way in which its data are used. Functioning as a tool for the evaluation of training, process-product relationships are used to test the appropriateness of the behaviors taught in training and to ascertain the degree to which the program can produce these behaviors in teachers. As in the relative gain model, "trained" and "untrained" groups of teachers are assessed to substantiate the expectation that training increases the teacher's use of target behaviors. Teachers are also observed prior to and after training for the same purpose. The information obtained in this manner can be used to gauge the importance of the behaviors emphasized in training (i.e., do they relate to pupil outcomes?) and also to determine the effectiveness of the training program in engendering these behaviors (i.e., does the trained teacher exhibit them more frequently than the untrained teacher?).

This information, however, is valuable not only to the trainer, but also to the researcher and the program developer who is interested in

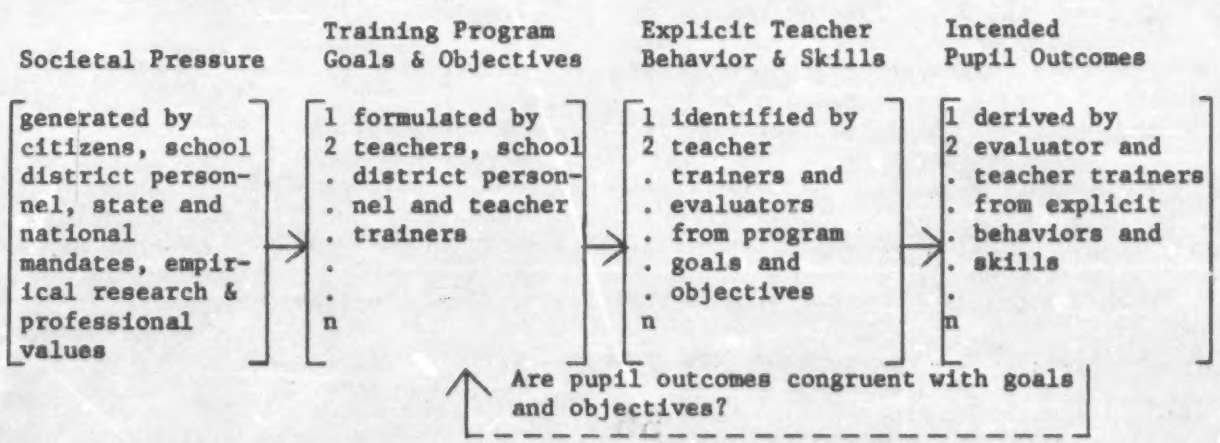
program revision. Traditionally, the process-product or teacher effectiveness researcher observes teacher process behaviors with a combination of high- and low-inference classroom observation instruments, incorporating a substantial number of teacher variables. The result is usually a large matrix of correlations among teacher process and pupil outcome variables, with no theory to explain the observed relationships. It is not uncommon to find a substantial number of these teacher behavior measures exhibiting low frequencies, indicating that the teacher had either no opportunity or no desire to use them on those occasions he or she was observed. The results of process-product research have sometimes been likened to the effects of a shotgun fired at long range: most of the shot completely misses the target, some barely misses it, and very little hits the bullseye. Without theory, however, the target itself is illusive, and the discovery of teacher-pupil relationships can be attributed as much to fortuitous probing as to systematic and informed investigation.

The process-product model employed in the context of a training program has the distinct advantage of having an identifiable target. The objectives and rationale of the program provide the framework in which process and, therefore product, behaviors are measured and correlated. While some variation is anticipated, the behaviors taught can be expected to match the objectives of the teacher and, thus, to occur with sufficient frequency during observation to provide stable variance estimates. Thus, the shotgun is exchanged for a rifle and the probability of a "hit" increased. While most process-product research is program-free, the efficiency of the model improves when it is used in the context of a specific training program with specific objectives. In this way, its evaluative and research functions are combined.

Following are the steps used in implementing the process-product model.

1. List Competencies. As in the two previous models, competencies must be identified from an examination of the training program or the objectives upon which the program is based. (Presumably these objectives were themselves derived from a list of competencies based upon the findings of process-product studies.) When findings from process-product studies provide the impetus for a training program, subsequent evaluation of the program can be considered an attempt at replicating the findings of the original studies. Even when objectives of the training program are not explicitly derived from process-product findings, the competencies taught may be operationalized in the same manner as in process-product investigations to allow a comparison of the two sets of results. Process-product findings similar to those listed on pages 23-28 in this report can be used for this purpose. Deriving competencies from a direct examination of program activities and materials is likely to be more time consuming but less risky: it increases the likelihood that behaviors taught but unspecified will be included in the evaluation and behaviors specified but clearly not taught will be eliminated.

2. Construct or Select Teacher and Pupil Measures. Developing instrumentation for the process-product model requires considerable circumspection and sensitivity on the part of the evaluator. It is the most vulnerable link in the chain that connects the content of the training program and the performance of pupils. And, as noted earlier, this connection is moderated by the trainee who must not only learn the content of the training program but also implement it sufficiently to affect pupil performance. In addition, the model must take into consideration those variables beyond the teacher's control which can weaken the impact of training. Thus, the instrumentation for measuring teacher behavior taught by the program and pupil behavior taught by the trainee must be sophisticated enough to sense and record the effect of the program on pupils. To increase the likelihood of developing valid instrumentation, the following model can be followed.



In this illustration, pupil outcomes are derived directly from explicit teacher behaviors and skills which, in turn, are derived from the goals and objectives of the training program. However, the link between program goals and intended pupil outcomes is strengthened if the two are congruent in the first place. Instrumentation should reflect a tight, overlapping relationship among program goals, teacher behaviors, and pupil outcomes. This congruence is especially critical if the process-product model is to successfully focus the impact of the training program on the performance of pupils.

Measures for assessing teacher behavior are usually selected from a class of instruments called "classroom observation scales." These instruments focus the observer's attention on either low-inference (i.e., discrete and specific) or high-inference (i.e., general and cumulative) behavior, and they take one of three forms: sign, category, or rating. A sign system records an event only once regardless of how often it occurs within a specified time. The behavior is given a code which indicates merely its presence or absence within a particular block of time. A category system, on the other hand, records a given teacher behavior each time it appears and, hence, provides a frequency count for the occurrence of specific behaviors, rather than a mere indication of their presence or absence. A frequency count may also be obtained using a modified sign system, called a rating instrument, which can be used to estimate the degree to which a particular behavior occurs.

Instrumentation for recording pupil outcomes can be either standardized or, preferably, criterion-referenced. As illustrated above, the link between teacher behavior and pupil outcome is as important as that between program content and teacher behavior. An adequate match between teacher behavior and pupil outcome can be achieved only if the instrument that records pupil performance is tailored to the explicit objectives of the teacher. The pupil performance that is measured must include no more and no less than that intended by the teacher. Criterion-referenced tests that are relatively brief and highly focused on program content fulfill this requirement. They can be prepared for each segment of the training program, and alternate forms constructed for pre- and posttesting.

3. Observe In Situ. Systematic classroom observation is the sine qua non of the process-product model--the characteristic that distinguishes it from the two previous models. The term "systematic" implies the rigorous application by two or more observers of sign, category, and/or rating systems in each teacher's classroom over randomly selected occasions. Generally, observation instruments, regardless of form, focus on either high- or low-inference behaviors. Those which ask an observer to judge the presence, absence, or degree of a teacher's warmth, organization, clarity, or enthusiasm, for example, require high inference, because item content does not specify discrete behaviors that must occur in order for a teacher to be considered warm, organized, clear or enthusiastic. Item content that requires

a cumulative judgment is considered high-inference. Observation instruments that specify exact behaviors to be recorded such as "teacher asks higher-order question" or "teacher uses example," require little inference on the observer's part. Low-inference item content generally reflects separate and distinct units of behavior that are easy to observe. The choice of scale type (sign, category, or rating) and item content (high or low inference) is determined by the nature of the behavior being measured. Low-inference category systems are generally most appropriate when specific, discrete and context-related behaviors are taught by the training program, and high-inference sign or rating systems are preferable when general, cumulative, and context-free behaviors are taught.

A second concern of the evaluator is the consistency or agreement, (i.e., reliability) between two independent observations recorded on the same coding instrument. In order to determine that the behavior in question can be observed and recorded with some precision, the reliability of the coding system must be established by correlating observations recorded by different raters using the same instrument and observing a teacher for the same period of time. Thus, for at least part of the classroom observation, two or more coders must be used.

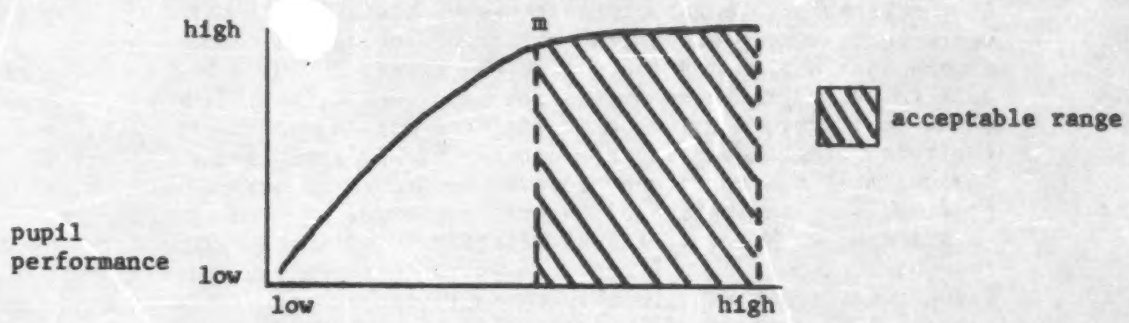
A second type of reliability, called generalizability, considers the number of occasions on which the teacher must be observed in order for the results obtained to generalize across all occasions on which the teacher could be observed. Indices of generalizability indicate whether the number of observation occasions selected is sufficient to study each behavior and whether the behaviors trained are sufficiently stable over a reasonable number of occasions and raters to be used as correlates of pupil performance. The ultimate purpose of all in situ observation is to quantify the extent to which the teacher has implemented the behaviors taught by the training program.

4. Measure Pre and Post Pupil Performance. As in the relative gain model, pre and post pupil performance must be measured to control the entry level behavior of pupils. Additional variables that are unrelated to the instruction provided by the teacher, but which can influence pupil outcome (such as aptitude, SES, and contextual variables, which may differ in non-random ways across classrooms, schools, or school districts) must be taken into account. Thus, process-product relationships represent findings obtained after appropriate statistical adjustments have been made.

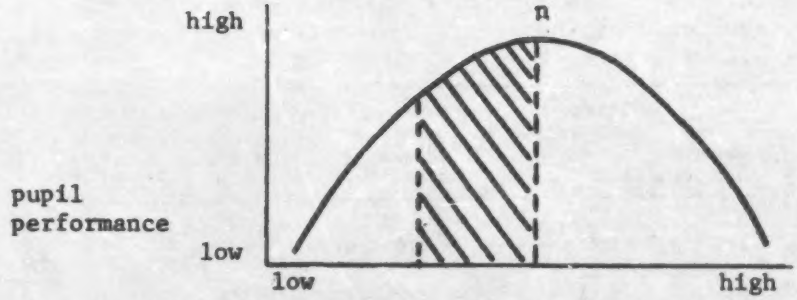
These adjustments can be made by computing residual gain scores or using covariance procedures. Residual gain is computed by correlating the pre- and posttest scores of all pupils, predicting a posttest score for each pupil on the basis of his or her pretest score, and subtracting this from the pupil's actual posttest score. This procedure creates a measure of gain which is independent of the pupil's initial standing and, therefore, more representative of the true change that has occurred during the observation. Analysis of covariance, which can be used to statistically control both the effect of pretest scores and other entry level variables, represents a somewhat more efficient procedure for accomplishing the same end.

5. Plot Relationships Between Teacher Process and Pupil Product Measures. Process-product correlations are an essential element in construction of teacher competencies. In the needs assessment and relative gain models, teacher competencies are inferred from teacher variables; only in the process-product model is the derivation of a competency empirically based. In any case, the formation of competencies should include specification of the level at which the teacher should perform a given behavior in order to be recognized as "competent." A competency is defined in terms of the level of proficiency that engenders meaningful pupil performance. The validation of various proficiency levels against meaningful classroom change is the primary contribution of the process-product model to competency development.

From correlations between teacher behavior (measured by classroom observation systems), and pupil outcome (measured by criterion-referenced tests), optimal levels of teacher behavior are determined. These relationships may take the following forms.



TEACHER BEHAVIOR A



TEACHER BEHAVIOR B

The relationship between teacher behavior A and pupil performance reaches asymptote at point m: application of the behavior at a level greater than m nets the learner little or no improvement in criterion behavior.

Thus, for teacher behavior A, the optimal level of proficiency is m and in order to have attained the competency, the teacher must exhibit level m at the completion of training.

The relationship between teacher behavior B and pupil performance is markedly curvilinear, failing to reach asymptote at any point; implementation of the behavior at a level greater than n will produce a decrement in pupil outcome. Thus, for teacher behavior B, the maximum acceptable level is n.

Relationships between process-product variables may, of course, take many other forms. But, regardless of form, they indicate to program developers which behaviors should be deleted from the training program (for lack of relationship to pupil outcome) and which levels of proficiency are most productive for each teacher behavior.

6. Construct Summary Tables. The final step in implementing the process-product model is the construction of tables summarizing the data obtained. These tables are used to compare either pre- and posttraining implementation of the behaviors taught or their use by trained and untrained teachers. They represent the most significant contribution of the process-product model to program evaluation. By indicating the extent to which the training program was able to engender the behaviors intended, and at what levels of proficiency, these tables provide the basis for judgments about the effectiveness of the total program and its components. Thus, while preceding steps investigated and confirmed process-product correlations, this step provides some indication of the program's worth. Process-product relationships, however, figure significantly in the selection of variables for which summary data are displayed. Only those variables that have shown significant relationship to pupil outcome are included on summary tables. The following examples illustrate some of the methods that can be used to display and summarize these data.

As in the needs assessment model, it is convenient to first divide competencies into knowledge, performance, and consequence categories. The evaluation of the training program should reflect trainee behavior on all three dimensions for each competency. This provides program developers with data about the type of competency as well as the degree of proficiency engendered by the program. A record of this progress might take the following form.

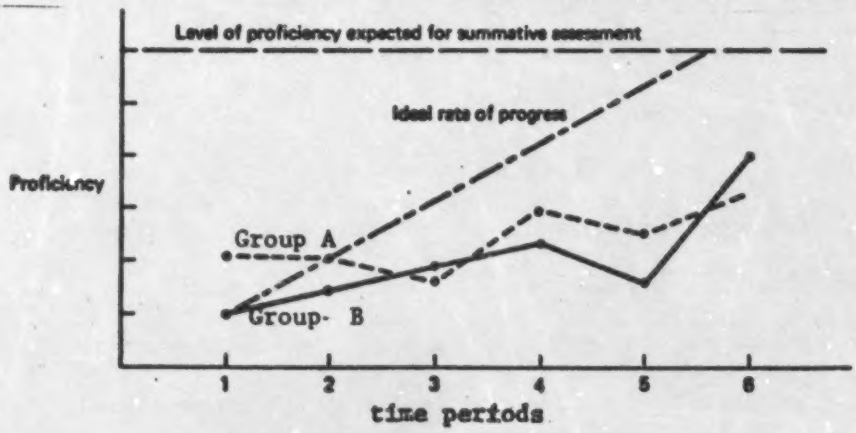
Competency areas	*Behaviors and skills*	*Showed knowledge of behavior or skill*	*Observed performing behavior or skill*	*Obtained desired (pupil) consequence as a result of using behavior or skill*
Questioning skills	1	• 9/7 ———	• 11/8	•
Modeling behavior	2	• 9/16 ———	• 10/21 ———	• 1/2
Reinforcement techniques	3	•	•	•
Individualization procedures	4	• 10/26 ———	• 10/14	•
	•			
	•			
	•			
	n	•	•	•

If frequency, percentage and variability data have been collected, the occurrence of selected behaviors during observation periods can be summarized in the following manner.

Behavior or skill observed	Frequency observed during most recent appraisal	Percent observed out of total or a selected subset of behaviors during most recent appraisal	Cumulative frequency over all previous appraisal periods	Cumulative percent over all previous appraisal periods	Mean frequency	Variation in frequency (standard deviation)
1						
2						
•						
•						
•						
n						

By using data from the table of frequencies and percentages, the evaluator can provide a continuous profile of teaching behaviors and skills for groups of teachers. A continuous profile allows the evaluator to summarize teacher performance in a number of areas and to illustrate the precise rate of trainee improvement across observation periods for each behavior or skill. It can provide a graphic indication of the proficiency level obtained by a particular teacher. This level can then be compared with the performance of various sub-groups of teachers.

For example, profiles can be constructed to compare performance of a group of trainees on selected behaviors and skills with: (a) that of other teachers in the school district; (b) the average performance of teachers who have previously participated in the training program; (c) an ideal performance profile, representing standards of proficiency suggested by the nature of the relationship between teacher behavior and pupil outcome, as determined in step (5) of the process-product model. A continuous profile of this type is indicated below.



The process-product model is perhaps best viewed as the final component in a sequence of evaluation activities that also includes the needs assessment and relative gain models. Theoretically, a training program can enter this sequence at any point, but this seldom occurs in practice. Due partially to its quantitative nature, the process-product model is often preceded by the relative gain and needs assessment models. Two additional factors, however, dictate its terminal position in the sequence. The first concerns the "press" typically experienced by a new program for immediate evaluation results. This press makes the needs assessment and relative gain models, which require less extensive instrumentation and little or no in situ observation, more attractive choices for the initial evaluation. The drain on human and fiscal resources that accompanies the process-product model may discourage relatively small training programs with limited funds and/or personnel, especially when the information needed may be adequately provided by the subjective responses of teachers.

Another reason for delaying process-product evaluation is the risk of negative or "null" findings. This risk is clearly greater when the method of evaluation assesses not only the effect of the training program on pupils, but also the link between this effect and specific teacher behaviors taught by that program. Programs that have already experienced some success with other, more descriptive models are more inclined to use the process-product approach and thereby risk exposure of their shortcomings. If the process-product relationships that form the basis of the training program are not confirmed, funding may be discontinued. Thus, the process-product model is most often employed when data from other sources have affirmed the efficacy of the training model and when revision of the program is a realistic option.

The amount of revision that can be tolerated and the extent to which training objectives can be operationalized can indicate the evaluation model which should be chosen. If the training staff is concerned with adjusting the program to correct minor problems or to assure that the quality of training remains constant over time and setting, the decisions they face are considered restorative in nature. The purpose of restorative evaluation is essentially to maintain the status quo and to provide a method of quality control. Thus restorative evaluation need only specify competencies in general terms and elicit from the trainee subjective responses about their value and attainment. Behavioral observation and measurement, generally, are not undertaken when the evaluation is intended only to check for minor variations in planned outcomes. In this case, the needs assessment model is the most appropriate.

If, however, the training staff wishes to examine larger problems in training format or content and if pupil outcome can be measured with appropriate instrumentation, the relative gain model is indicated. Because it produces relatively detailed pupil data, the relative gain model will likely uncover more extensive weaknesses and suggest greater revision in the program than would the needs assessment approach.

Finally, if the training staff wishes to make decisions about the values and assumptions underlying the program and can operationalize the teacher process and pupil outcome behaviors, the process-product model is the best choice. However, in this case behavioral objectives for the program must be stated in reference to both teacher and pupil, and the anticipated relationship between the two must be expressed. Because the process-product model examines the theory on which training is based, it represents a risk to the training program. If the direction

and magnitude of relationships predicted between teacher and pupil behavior are not confirmed, the validity of the program may be questioned. Generally, the degree of revision indicated by process-product data is greater than that suggested by information obtained from the other two models.

Like the relative gain model, the process-product approach is viewed with some anxiety by the teacher. Not only is pupil behavior measured, as before, but also complex variables which may influence the behavior of pupils in spite of, not because of, teacher performance. Thus, it is hard to escape the possibility that negative findings will be ascribed to teacher behavior when the contextual or antecedent conditions virtually assure lack of pupil growth. This threat and anxiety can be minimized, however, if teachers are involved in the selection of both posttest content and contextual and pupil variables to be statistically controlled.

Whereas the relative gain model is most frequently employed by school districts engaged in short-term training, the process-product model is most often chosen by the school, college, or department of education offering degree or certification programs. A training program must be of considerable duration in order to justify the time and resources required by the process-product model. Also, the combined research and evaluative functions of the process-product model often appeal to university training staff whose interests are commonly divided between research and evaluation. And, to consider a practical issue, process-product studies can be classified as research or evaluation, depending on the funding source available. In any case, the process-product model serves two purposes equally well by relating teacher process to pupil product, thus testing the predictions theorized, and by providing evaluative data that describe

the implementation of teacher behaviors in the inservice classroom. As noted, process-product studies conducted for a specific program may actually represent a more appropriate research paradigm than program-free studies, since the former are more likely to clarify both the theory and the nature of the process-product relationships it predicts.

The definition of evaluation implicit in the process-product model goes beyond the simple normative improvement of pupils and the subjective judgments of trainees about "what is" and "what should be," to determine the actual worth of the program. The model's capacity to test the theory employed in selecting training content and to assess the ultimate target of training permit a true assessment of the program's value. The theory that links program content to pupil outcome is what the model tests, and since all program content derives from the theory, refutation of the theory implies refutation of the program. According to our definition, program worth is a function of pupil performance and its relationship to teacher behavior. When process-product relationships are strong, the program is worthwhile. When relationships are weak, non-existent, or negative, the program has little or no value.

Conclusion: Applications and Summary

The concluding portion of this paper discusses the various contexts in which the needs assessment, relative gain, and process-product models can be used. Because one of the most salient problems faced by training staff is selection of the best model for a particular setting, case studies are presented to illustrate the contextual characteristics most often associated with each of the evaluation models described in this paper. Key characteristics or variables that suggest the use of one model over another are also identified. These descriptions necessarily include generalizations about the advantages and limitations of each model for different types of settings and therefore should be considered suggestive rather than definitive.

Case 1

A relatively new college of education has just initiated a master's degree program for experienced elementary school teachers. The program is based on standard curriculum concepts and general principles of education. One unique aspect of the program, however, is the requirement that degree candidates take at least 1 year of formal course instruction at the training institution and then spend an additional year as inservice teachers, applying certain teaching competencies in the classroom, under the regular observation of a member of the training staff. The 2nd year of the program is nearing completion and its first class is about to graduate when the dean of the college of education requests evidence from the training staff that the program is fulfilling the needs of teachers and the communities they serve. Ostensibly, the evaluation is intended to suggest program revisions that can be made before the next training cycle,

but the staff suspects that a "tight money" year is causing the dean to consider terminating the program. Five hundred dollars and a 1/2-time graduate student have been allocated to the evaluation for 8 weeks, at the end of which time a report is to be completed.

Suggested Model: Needs Assessment.

Distinguishing Characteristics

1. Only small adjustments are acceptable if the program is to be continued for another cycle. Large-scale revision or complete reconceptualization would seem to preclude continuation of the program and, in turn, further evaluation and revision.

2. Because the training curriculum is quite general, it would be difficult or even impossible to operationalize selected process and product variables without an extensive examination of that curriculum. It is unlikely that the required time and resources would be made avaliable for this purpose.

3. Limited time and personnel preclude *in situ* observation and instrument construction, though a broad survey measure might be created. Avaliable funds might best be spent for questionnaire duplication and a mailing to recent graduates of the program.

4. Because evaluation results are to be used for immediate decision making rather than examination of the theory or concepts upon which the program is based, descriptive data that capture the impressions of recent graduates are probably most useful.

5. Because there are no data affirming the efficacy of the program, some positive information may be needed, along with suggestions for revision, to continue the training.

Case 2

A school district serving a metropolitan, low-SES area has for the past 6 years, offered a course in behavior management to its secondary school teachers. In the original proposal the program's expressed purpose was "to increase math and reading achievement through a reduction

of classroom discipline problems." The course consists of 10 2-hour sessions, 1 per week for which university credit is given. Because the school district has allotted a very small proportion of the budget to instructional development, the training staff has decided to examine the format and content of their training sessions with an eye toward possible revision. Since the course is taught only during the fall, the staff has designated the final 6 weeks of the current training cycle and the first 6 weeks immediately following as the data collection period.

Suggested Model: Relative Gain (or if funds permit, Process-Product).

Distinguishing Characteristics

1. Moderate to considerable revision seems acceptable to the training staff. Objectives of the evaluation seem to involve both revision and affirmation of training content and format.

2. Due to the narrow content of the training program, i.e., behavior management, pupil and possibly teacher outcomes could be operationalized from existing program descriptions, or even from training materials, without incurring too great an expense.

3. Time is apparently not a factor. However, personnel to train observers, to serve as observers, and to construct process instrumentation may be limited. The staff should consider the availability of student observers, the possibility of using existing rather than new process instrumentation, and the complexity (in terms of observer training) of the process behaviors to be recorded.

4. The program's expressed purpose implies a process-product link. Thus, an examination of the theory that poses this link should concern the training staff as much as constructive evaluation of program content. The training staff might be asked to address this question in order to determine the need to examine the theoretical underpinnings of the program.

5. The program has apparently been conducted for some time without criticism. Thus, its continuation seems likely even in the face of needed revision. In light of this, the training staff seems willing to accept a more stringent accountability model.

Case 3

Four schools have been selected by the evaluation department of a large school district to field test a new inservice curriculum package for teaching reading in the elementary grades. This package, developed by a national research and development institute, is being implemented in the school system for the first time. Its design is purportedly based on concepts of imagery and word association as described in a recently published cognitive theory of learning. The authors of the theory claim that among this curriculum's benefits is a substantial increase in the reading comprehension of minority children who have been taught by teachers using the prescribed methods. Because the curriculum appeared promising in a earlier product evaluation conducted by the institute that developed it, the school district has decided to fund another evaluation at a relatively high level for a 1 year period. The school district's primary intent is to determine whether the curriculum package can actually train teachers in the specified strategies with the effects claimed, and, hence, be disseminated to all schools in the district the following year.

Suggested Model: Process-Product.

Distinguishing Characteristics

1. Refutation of the theory and, thus, major changes in the curriculum would seem an acceptable outcome of the study. The school district is, of course, concerned primarily with the effect of the curriculum on the achievement of its pupils, and this end seems best served by an examination of the relationships between teacher behavior and pupil outcome predicted by the theory.

2. Due to the explicit identification of the theory on which the program is based, intended process and product variables should be easy to operationalize by consulting the theory. Here, existing documentation may be sufficient to specify key process and product variables in terms of measurement procedures.

3. Both time and technical staff seem adequate for an extensive process-product evaluation. While some adjustments may be necessary, there will probably be no need to revert to a less comprehensive model.

4. Data will be used both to confirm the theory and to describe the program's effects on teachers. These data, of course, will also interest the research and development institute that has devised the curriculum.

5. Because the previous appraisal of the curriculum was limited to pupil outcomes, a more comprehensive evaluation is warranted if earlier findings are to be confirmed or enhanced.

Case 4

The special education department of a college of education has recently implemented a series of Saturday morning workshops designed to train inservice elementary teachers in a variety of techniques for teaching the handicapped child in the regular classroom. The program consists of four 2-hour workshops, each conducted by a different instructor in a lecture-discussion format. The purpose of these sessions is to make regular inservice teachers aware of the different teaching strategies promoted by each of four instructors who are presumably experts in their respective fields. The program is funded by a federal grant through the special education department. While funds have not been provided for an evaluation of these sessions, the application for renewal of program funding clearly requires evidence supporting the "success" of the first series of workshops. The department chairman has decided to use discretionary funds from his departmental budget to fund an evaluation in the amount of 5% of the cost of this $2,500 program.

Suggested Model: Needs Assessment.

Distinguishing Characterics

1. Minor revisions in content and/or speakers could be made, but funds, and to some extent objectives and allotted time, seem to limit the program to its present format.

2. While some process variables could be specified, it is unlikely that pupil outcome could be identified at a sufficient level of detail to be attributed to the program. The primary intent of the program is to create "awareness," which seems realistic given the time and resources devoted to it. Implementation of the techniques taught might or might not occur, but in either case, would be difficult to attribute to the program, per se.

3. Time and personnel permit no more than a questionnaire survey of program participants. The cost of questionnaire duplication and postage would probably just about match the amount of funds available.

4. The purpose of this evaluation is clearly descriptive. There is no implication that the workshops are linked by a theory, or the techniques taught linked to pupil outcome. Specification of pupil outcome, even in broad terms, would be difficult.

5. The lack of previous evaluation data and the limited resources available for the current evaluation suggest that this initial effort should remain small and manageable, permitting only minor adjustments in scope and format.

Case 5

A school district has decided to base an inservice training project on the findings of a large-scale, nationwide process-product study completed the previous year. In this study, the process variable relating most significantly to the achievement of elementary pupils was the teacher's question-asking behavior (e.g., whether the question was higher order, lower order, affective, cognitive, process, or substantive). Because the school system's most pressing concern was the relatively low achievement

of its secondary students, district personnel decided to develop a 7-week inservice training program to teach these questioning skills to its secondary school teachers. Since these process-product findings came from a program-free study, the staff responsible for developing training materials had to infer from these variables the nature of the materials needed. They also had to assume that such training would produce effects at the secondary school level. The project was considered somewhat risky by school district personnel since the initial process-product study had used elementary school children, and there was no guarantee that its findings would apply to older students. For this reason, the majority of funds were devoted to developing training materials, with the stipulation that additional funds would be available for more extensive evaluation and revision of materials if the initial evaluation were encouraging. Thus, the school district limited this initial evaluation to determining whether the earlier process-product findings might be generalized to the secondary school level.

Suggested Model: Relative Gain.

Distinguishing Characteristics

1. Realizing the risk involved, the school district is apparently willing to accept either an entirely positive or an entirely negative result. Thus, tolerance for change seems high, since the school district is willing to discontinue the program should initial results be discouraging.

2. While both process and product behaviors seem easy to operationalize, funds may be sufficient to measure only pupil outcome. An examination of pupil achievement from a group of trained teachers, vis-a-vis pupil achievement from an untrained group, could produce the result desired. That is, if the effectiveness of the program was revealed in student achievement, a later study could ascertain the exact teacher behaviors that produced the difference between trained and untrained teachers.

3. While time may not be a factor, the demand on personnel may be great, especially if criterion-referenced achievement tests must be constructed. This can represent a considerable investment in resources, leaving little for systematic observation of the degree to which the trained teachers are implementing the questioning skills.

4. While the original process-product findings upon which the project was based must eventually be replicated in this new context, such relationships can be inferred from differences in the achievement of pupils of trained and untrained teachers. Thus, if comparison groups are used, much of the information yielded by the process-product model would be incorporated in this "control" and "experimental" version of the relative gain approach.

5. While there are no evaluative data on project materials themselves, the objectives of the study seem to require that the impact of the materials on pupils be measured. Given the findings of the original process-product study, teacher perceptions about the training program seem insufficient data upon which to decide the usefulness of subsequent and more extensive evaluations.

REFERENCES

Borich, G. The appraisal of teaching: Concepts and process. Reading, MA.: Addison-Wesley, 1977.

Brophy, J. and Evertson, C. Process-product correlations in the Texas teacher effectiveness study: Final report (Res. Rep. 74-4). Austin, Texas: Research and Development Center for Teacher Education, 1974.

Brophy, J. and Good, T. The Brophy-Good dyadic interaction system. In A. Simon & E. Boyer (Eds.), Mirrors for behavior: An anthology of observation instruments continued, 1970 supplement (Vol. A). Philadelphia: Research for Better Schools, Inc., 1970.

Good, T.L. and Grouws, D.A. Process-product relationships in 4th grade mathematics classes. Columbia, Missouri: College of Education, University of Missouri, 1975.

McDonald, F.J., Elias, P., Stone, M., Wheeler, P., Lambert, N., Calfee, R., Sandoval. J., Ekstrom, R., and Lockheed, M. Final report on phase II beginning teacher evaluation study. California Commission on Teacher Preparation and Licensing, Sacramento, California. Princeton: Educational Testing Service, 1975.

Peer, G. and Pugues, W. A national survey of teacher education follow-up practices. Paper presented at the annual meeting of the American Association for Colleges of Teacher Education, Chicago, Illinois, February, 1978.

Rosenshine, B. Teacher behaviors and student achievement. Windsor, Berks, England: National Foundation for Educational Research in England and Wales, 1971.

Ryans, D.G., Characteristics of teachers. Washington, D.C.: American Council on Education, 1960.

Scriven, M. The methodology of evaluation. In AERA monograph series on curriculum evaluation, Vol. 1. Chicago, Illinois: Rand-McNally, 1967.

Simon, A. and Boyer, B., Eds., Mirrors for behavior: An anthology of observation instruments. Philadelphia: Research for Better Schools, 1970.

Soar, R.S. An integrative approach to classroom learning. Philadelphia, Temple University, 1966.

Soar, R.S. and Soar, R.N. An empirical analysis of selected follow-through programs: An example of a process approach to evaluation. In I.J. Gordon (Ed.), Early childhood education. Chicago: National Society for the Study of Education, 1972.

Stake, R.E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Stallings, J. and Kaskowitz, D. Follow-through classroom observation evaluation, 1972-73. Menlo Park, California: Stanford Research Institute, 1974.

Stufflebeam, D., Foley, W., Gephart, W., Guba, E., Hammond, R., Merriman, H., Provus, M. Educational evaluation and decision making. Itasca, Illinois: F.E. Peacock, 1971.